Universidad
Tecnológica
de Pereira

Metodología de Visualización de Datos Utilizando Métodos Espectrales y Basados en Divergencias para la Reducción Interactiva de la Dimensión

Andrés Javier Anaya Isaza

Tesis de Maestría

Universidad Tecnológica de Pereira
Facultad de Ingenierías: Eléctrica, Electrónica, Física y Ciencias de la Computación
Pereira - Risaralda, Colombia
2017

Metodología de Visualización de Datos Utilizando Métodos Espectrales y Basados en Divergencias para la Reducción Interactiva de la Dimensión

Andrés Javier Anaya Isaza

Tesis presentada como requisito parcial para optar por el título de:
Magíster en Ingeniería de Sistemas y Computación

Director:
PhD. Diego Hernán Peluffo-Ordóñez

Co-director:
PhD. Jorge Iván Ríos Patiño

Universidad Tecnológica de Pereira
Facultad de Ingenierías: Eléctrica, Electrónica, Física y Ciencias de la Computación
Pereira - Risaralda, Colombia
2017

Data visualization methodology using spectral and divergence-based methods for interactive dimensionality reduction

Andrés Javier Anaya Isaza

Thesis presented as a partial requirement for the degree of:
Magíster en Ingeniería de Sistemas y Computación

Advisor:
PhD. Diego Hernán Peluffo-Ordóñez

Co-advisor:
PhD. Jorge Iván Ríos Patiño

Universidad Tecnológica de Pereira
Facultad de Ingenierías: Eléctrica, Electrónica, Física y Ciencias de la Computación
Pereira - Risaralda, Colombia
2017

(Dedicatoria)

A mi madre y mis ángeles.

Porque hicieron que ante la inmensidad de la oscuridad, existiera un hermoso paraíso a través de la luz de sus sonrisas, inspirando mi vida para navegar en este mar de contrastes, sin perder el rumbo de la esperanza.

Andrés Javier Anaya Isaza

Agradecimientos

A Dios, por darme fuerza para no desfallecer y abandonar este duro camino de aprendizaje lleno de sacrificios, el cual encontré gente maravillosa que me enseñó el sendero de la sabiduría a través del conocimiento compartido.

A mi bella Madre, que siempre estuvo conmigo y cree ciegamente en mis capacidades, incentivando mis fortalezas para ser mejor cada día, a través de su ejemplo y sabias palabras. A mi Hermano le aprendí el significado de la perseverancia y disciplina que trazaron el sueño de la academia en mi vida, para ayudar y formar a las nuevas generaciones.

Mis hermosos ángeles que inspiraron y cambiaron mi vida para continuar con este camino y labrar un futuro juntos, porque el fruto de su amor es mi motor y motivación para ser cada día mejor persona.

Estoy especialmente agradecido con mi director de tesis PhD. Diego Peluffo Ordoñez por ser una persona tan especial y guiarme con extrema paciencia en este proceso enriquecedor, además de mostrarme el camino con claridad de los procesos de la ciencia y como buscar el sendero de la excelencia a través de su testimonio de vida y disciplina, nunca tendré como pagarte querido maestro.

Agradezco a mi co-director de tesis, PhD Jorge Iván Ríos por todo el apoyo a lo largo de este proceso y las enseñanzas en inteligencia artificial, sin tu ayuda todo hubiera sido diferente, gracias, maestro Jorge.

Agradezco a todos los miembros del equipo de DataVis por el apoyo y todo el conocimiento compartido a lo largo de este proceso y la motivación para seguir formándonos en todos estos campos de conocimiento infinito para compartir con todos aquellos que nos necesitan.

Resumen

Las tareas de reconocimiento de patrones aplican métodos que evolucionan de manera equivalente al crecimiento de los datos, alcanzando métricas eficientes en términos de optimización y rendimiento computacional aplicado a exploración, selección y representación de datos. No obstante, los resultados brindados por dichos métodos y herramientas podrían resultar ambiguos y/o abstractos para el usuario, haciendo que su aplicación sea compleja, aún más si no cuentan con un conocimiento previo de los datos. Tener un conocimiento a priori garantiza en el mayor de los casos la correcta selección del modelo, así como también algoritmos y métodos adecuados. Sin embargo en datos masivos, donde éste conocimiento es escaso y poco factible, los procesos de interpretación podrían ser arduos para los usuarios, especialmente, para aquellos usuarios no expertos.

En consecuencia, han surgido diversos problemas que debe enfrentar el reconocimiento de patrones, entre los más importantes se encuentran: La reducción de dimensión, la interacción con grandes volúmenes de información, la interpretación y la visualización de los datos. Lo anterior puede enmarcar conceptos de controlabilidad e interacción que son propiedades, en su mayoría, ausentes en las investigaciones típicas dentro del campo de reducción de dimensión.

Esta tesis presenta un nuevo enfoque de visualización de datos, basada en la mezcla interactiva de resultados de los métodos de reducción de dimensionalidad (RD). Tal mezcla es una suma ponderada, cuyos factores de ponderación son definidos por el usuario a través de una interfaz visual e intuitiva. Además, el espacio de representación de baja dimensión producida por métodos de (RD) se representan gráficamente mediante diagramas de dispersión alimentados a través de una visualización de datos interactiva controlada. Para ello, se calculan las distancias entre pares por similitud y se emplean para definir el gráfico a representar en el diagrama de dispersión. El enfoque de visualización permite al usuario combinar interactivamente métodos (RD) mientras se proporciona información sobre la estructura de datos original, haciendo la selección de un esquema de un (RD) más intuitivo.

Palabras clave

Reducción de la dimensión, Visualización de información, Analítica visual, Reconocimiento de patrones, Interacción Humano-Maquina.

Abstract

The methods applied to pattern recognition tasks evolve in parallel with the growth of the data, Achieving efficient metrics in terms of optimization and computational performance applied to exploration, selection and representation of data. However, the results provided by such methods and tools may be ambiguous and / or abstract to user, making their application complex, even more if they dont have prior knowledge of data. Having an a priori knowledge guarantees, in most cases, the correct selection of the model, as well as adequate algorithms and methods. However, in massive data, where this knowledge is scarce and unfeasible, the interpretation processes could be arduous for users, especially for those users who are not experts.

Consequently, several problems have arisen that must face the recognition of patterns, among the most important are: The dimensionality reduction, interaction with large volumes information, interpretation and visualization of data. The above may be encompassed by the concepts of controllability and interaction which are mostly absent in typical investigations within the field of dimensionality reduction.

This thesis presents a new interactive data visualization approach based on mixture of the outcomes of dimensionality reduction (DR) methods. Such a mixture is a weighted sum, whose weighting factors are defined by the user through a visual and intuitive interface. Additionally, the low-dimensional representation space produced by DR methods are graphically depicted using scatter plots powered via an interactive data-driven visualization. To do so, pairwise similarities are calculated and employed to define the graph to be drawn on the scatter plot. Our visualization approach enables the user to interactively combine DR methods while provided information about the structure of original data, making then the selection of a DR scheme more intuitive.

Keywords

Dimensionality Reduction, Information Visualization, Visual analytics,Pattern Recognition,Human-Computer Interaction..

Tabla de Contenido

Lista de Figuras

Lista de Tablas

1. Nomenclatura

En esta sección, se incluyen la notación utilizada en el desarrollo del documento, así como también los símbolos generales y abreviaturas utilizadas en las diferentes partes de la presente Tesis de Maestría. A continuación se expone la notación con sus correspondientes términos en la sección 1.1 y las abreviaturas en la sección 1.2.

1.1. Notación

Notación	Término
$\mathbb{N}$	El conjunto de números naturales positivos: $\{0, 1, 2, 3, ...\}$
$\mathbb{R}$	El conjunto de números reales
y, x	Variables aleatorias conocidas o desconocidas tomando sus valores en $\mathbb{R}$
$\mathbf{A}$	Una matriz
$a_{i,j}$	Una entrada de una matriz $\mathbf{A}$
	(Situado en el cruce de la $i$ésima fila y la $j$ésima columna)
N	Número de puntos en el conjunto de datos
M	Número de prototipos en el libro de códigos $\mathbf{C}$
D	Dimensión del espacio de datos (que normalmente es $\mathbb{R}^D$)
P	La dimensión del espacio latente (que es generalmente $\mathbb{R}^P$)
	(o su estimación como la dimensión intrínseca de los datos)
$\mathbf{I}_D$	Matriz D-dimensional de identidad
$\mathbf{I}_{P\times D}$	Matriz rectangular que contiene la primera P filas de $\mathbf{I}_D$
$\mathbf{1}_N$	Vector de columna N-dimensional, que contiene uno por todas partes
$\mathbf{y}$	Vector aleatorio en el espacio de datos conocido: $\mathbf{y} = [y_1, ..., y_d, ..., y_D]^T$
$\mathbf{x}$	Vector aleatorio en el espacio latente desconocido: $\mathbf{x} = [x_1, ..., x_p, ..., x_P]^T$
$\mathbf{y}(i)$	El vector i-ésimo del conjunto de datos
$\mathbf{x}(i)$	(Vector latente no conocido) que generó $\mathbf{y}(i)$
$\hat{\mathbf{x}}(i)$	La estimación de $\mathbf{x}(i)$
$\mathcal{Y}$	El conjunto de datos $\mathcal{Y} = \{..., \mathbf{y}(i), ...\}_{1\leq i\leq N}$
$\mathcal{X}$	El conjunto (desconocido) de vectores latentes que generaron $\mathcal{Y}$
$\hat{\mathcal{X}}$	La estimación de $\mathcal{X}$
$\mathbf{Y}$	El conjunto de datos en notación matricial: $\mathcal{Y} = [..., \mathbf{y}(i), ...]_{1\leq i\leq N}$
$\mathbf{X}$	El conjunto (desconocido) ordenado de vectores latentes que generaron $\mathbf{Y}$

Abreviatura	Término
$\hat{\mathbf{X}}$	Estimación de $\mathbf{X}$
$\mathcal{M}$	Un colector (indicado como un juego)
$\mathbf{m}$	La notación funcional de $\mathcal{M} : \mathbf{y} = \mathbf{m}(\mathbf{x})$
$E_x\{x\}$	La expectativa de la variable aleatoria x
$\mu_x(x)$	El valor medio de la variable aleatoria x
	(Computo con sus valores conocidos $x(i), i = 1, ..., N$)
μ_i	El momento centrado en el orden i-ésimo
μ'_i	El i-ésimo-orden del momento aproximado
$\mathbf{C_{xy}}$	La matriz de covarianza entre los vectores aleatorios $\mathbf{x}$ y $\mathbf{y}$
$\hat{\mathbf{C}}_{\mathbf{xy}}$	La estimación de la matriz de covarianza
$f(\mathbf{x}), \mathbf{f}(\mathbf{x})$	Función unidireccional o multivariable del vector aleatorio $\mathbf{x}$
$\langle \mathbf{y}(i) \cdot \mathbf{y}(j) \rangle$	Producto escalar entre los dos vectores $\mathbf{y}(i)$ y $\mathbf{y}(j)$
$d(\mathbf{y}(i), \mathbf{y}(j))$	Función de distancia entre los dos vectores $\mathbf{y}(i)$ y $\mathbf{y}(j)$
	(A menudo una distancia espacial, como la euclidiana)
	acortado como $d_{\mathbf{y}}(i, j)$ o $d_{\mathbf{y}}$ cuando el contexto es claro
$\delta(\mathbf{y}(i), \mathbf{y}(j))$	Distancia geodésica o grafo entre $\mathbf{y}(i)$ y $\mathbf{y}(j)$

1.2. Abreviaturas

Abreviatura	Término
RD	Reducción de la Dimensión
LDR	Reducción de la Dimensión Lineal
$NLDR$	Reducción de la Dimensión No-Lineal
SVD	Descomposición de Valores Singulares
EVD	Descomposición de Valores Propios
CCA	Análisis de Componentes Curvilíneos
CDA	Análisis de Distancias Curvilíneas
GTM	Mapeo Generativo Topográfico
$HLLE$	Hessiana de LLE, (ver LLE)
LE	Mapas Laplacianos
LLE	Embebimiento Lineal Local
MDS	Escalamiento Multidimensional
NLM	(Sammon's) Mapeo No-Lineal
PCA	Análisis de Componentes Principales
SDE	Embebimiento Semidefinido
SNE	Embebimiento Estocástico Vecinal
SOM	Mapas Auto-Organizados

Parte I.

Preliminares

2. Introducción

Eric Schmidt, Executive Chairman, Google "Desde los inicios de la civilización hasta el año 2003, la humanidad generó cinco exabytes 10^{18} ***bytes de datos. Ahora se produce cinco exabytes cada dos días y el ritmo se está acelerando" Agosto 2010.***

Los seres humanos crean y almacenan constantemente información en cantidades astronómicas. Haciendo una relación de bits y bytes con datos del último año y posteriormente almacenando esta información en CD's, se podría crear una torre desde la Tierra a la Luna y de regreso [2]. Esta contribución a la acumulación masiva de datos se puede encontrar en varias industrias. Las empresas mantienen grandes cantidades de datos transaccionales, recopilando información sobre sus clientes, proveedores, operaciones, etc. Del mismo modo ocurre en el sector público [3]. En países de Europa, Asia, América del Norte, Centro, Sur entre otros, se evidencia el proceso de gestión de la información en bases de datos relacionales y no relacionales, en su mayoría contiene datos tales como: censo de población, registros médicos, impuestos, etc. A esto, se suma las transacciones financieras realizadas en línea desde dispositivos móviles, el análisis de redes sociales como las afamadas Twitter y Facebook (En Twitter cerca de 12 Terabytes de Tweets son creados diariamente así como también, Facebook almacena alrededor de 100 Petabytes de fotos y videos [4], así mismo los dispositivos IoT(Internet of Things) a través de sus sensores, actuadores, y demás periféricos de entrada y salida, generan diversas salidas como registros de ubicación geográfica por coordenadas GPS entre otras. En términos del común, todas aquellas actividades rutinarias con un smartphone generan un promedio de 2.5 quintillones diarios De bytes en el mundo [2] [3].

1 quintillon = 10^{30} = 1,000,000,000,000,000,000,000,000,000,000

Acorde al reporte cisco vini-2011 [2], cerca del 2011 y 2016, la cantidad de tráfico móvil de datos crecerá a razón de cambio anual de 78%, Así como el número de dispositivos móviles conectados a Internet, excederá el número en habitantes del planeta. Según la organización de las Naciones Unidas, la población mundial alcanzará los 7.500 millones en 2016, de tal manera que habrá cerca de 18.900 millones de dispositivos conectados a la red en todo el mundo, esto conlleva a un tráfico global de datos móviles de 10.8 Exabytes al mes o 130 Exabytes al año. Este volumen de tráfico en 2016 fue equivalente a 33 mil millones de DVDs anuales o 813 cuatrillones de mensajes de texto [2].

No sólo, los seres humanos contribuyen al enorme crecimiento de la información, sino la comu-

nicación llamada máquina-a-máquina (M2M) cuyo valor en la creación de datos masivos es muy importante [5] [6]. Un ejemplo de ello sucede cuando sensores digitales se instalan en contenedores para determinar la ruta generada durante la entrega del paquete y esta información se envía a las empresas de transporte, otro caso de estudio hace referencia a sensores de sensores eléctricos, para determinar la energía de consumo a intervalos regulares de tal modo que ésta información se envíe al sector empresarial energético. Se estima que, más de 30 millones de sensores están interconectados en diferentes sectores como: automoción, transporte, industria, servicios comerciales, etc. Se espera que este número aumente un 30% anual [3] [7].

Esta tesis, enmarca el diseño de una metodología de Análisis de Datos Visuales Utilizando Métodos Espectrales para la Reducción de Dimensión Interactiva. El cual brinda la posibilidad de interactuar con el usuario en los métodos de selección y reducción de la dimensión de la mezcla. También se diseñó no sólo el modelo de interacción y la interfaz, sino también la formulación de un método generalizado de reducción de la dimensión que permite la selección intuitiva y los métodos de mezcla. Un aspecto transversal en todas las etapas del diseño de la metodología de análisis visual tiene una relación con el coste computacional que impide que el objetivo de esta metodología sea realmente interactivo, es decir, en tiempo real. En esta tesis se realizaron todos los diseños e implementaciones de prueba en ambientes de bajo coste computacional.

2.1. Planteamiento del Problema

Los problemas que surgen son: En términos de percepción humana, cuando la representación de datos no se ajusta a los parámetros de fácil interpretación, se determina que carece de; buen nivel de abstracción o simplemente posee varias capas de procesamiento, el cual hace compleja su comprensión. A este nivel, es indispensable desarrollar herramientas que permitan la percepción holística del problema a resolver, con la finalidad de operar sobre el medio ambiente circundante y observar las propiedades del entorno de una manera directa e indirecta, de esta manera el proceso ejecuta acciones que resultan relevantes bajo el marco experimental del entorno real. Desde el punto de vista del diseño de interfaces de usuario, se desea conocer que control (botón) virtual puede ser accionado por medio de un ratón, un cursor, o aún, otro botón. Esto no es una interacción directa con el mundo físico [8][9]. Desde otra perspectiva, si se utiliza entornos y reglas comunes como las de percepción del mundo real y como se interactúa con él, los resultados son más significativos, en términos de entendimiento. Por ejemplo; si se utiliza la regla de las perspectivas, se sabe que un objeto a mayor distancia se mira más pequeño, y sobre esta regla nuestro cerebro se condiciona para abordar una acción. Las interfaces pueden ser eficientemente comprendidas por cualquier persona sin importar de qué cultura procede, si se diseñan en torno a las teorías de la percepción. Además, las acciones tradicionales de análisis de datos implican la ejecución de un único algoritmo o técnica a lo largo de todo el proceso, sin embargo, las características ausentes

en un método podrían encontrarse en otros, y por tanto, una apropiada integración de ellos lograría potenciar sus propiedades y generar una mejor representación de los datos. [10]

Esta tesis de maestría busca solventar las dificultades mencionadas, siendo un puente entre el dominio de dos contextos de investigación, como son la Reducción de la Dimensionalidad (RD) e Visualización de la Información (INFOVIZ (IV)), dos campos que hacen parte del Aprendizaje de maquina (Machine learning), específicamente de Minería de datos (Data Mining) y Reconocimiento de patrones (Pattern recognition) que se refieren respectivamente a la representación y visualización de información cuantitativa multivariada, especialmente con un número significativamente grande de variables. Esto se puede hacer importando los conceptos de controlabilidad e interacción que están en el dominio de la (IV) y proyectándolos al (RD) para hacer un método de reconocimiento de patrones (pattern recognition) controlable e interactivo, ya que el objetivo de la (IV), es desarrollar métodos gráficos que presenten la información más relevante para el usuario, bajo criterios de controlabilidad, donde el usuario pueda decidir cuál es el mejor modo de representar la información subyacente de sus datos en base a su objetivo de análisis, utilizando una interfaz que responda rápidamente a los cambios de parámetros, es decir, utilizar las propiedades de la visualización para hacer más legibles los resultados de la reducción de dimensionalidad, así como más cercanos al usuario a través de combinaciones de diversos métodos de manera interactiva y amigable, de tal forma que permita la consecución gradual del objetivo en donde los pasos intermedios sean abordados en base a las teorías de la percepción humana, dando lugar a nuevos diseños de interfaces que permitan: Operaciones mentales con un rápido acceso a grandes cantidades de datos fuera de la mente, inferencia cognitiva, reducción de la demanda de la memoria de trabajo y co-participación de la maquina en una tarea conjunta, mediante el cambio gradual de las visualizaciones de forma dinámica [10].

Uno de los factores más importantes del método propuesto es la interactividad síncrona que permitirá que los métodos (RD) se ajusten de acuerdo al criterio del usuario, quien aún sin conocer específicamente los métodos que se han aplicado, podrá obtener resultados confiables. Esta tesis podría representar un aporte en el área de Aprendizaje de máquina (Machine learning), y Reconocimiento de patrones (Pattern recognition) en términos de realizar una visualización eficiente permitiendo a un usuario, no experto o sin previo conocimiento de los métodos, obtener resultados visuales de fácil interpretación mediante el uso de una interfaz interactiva de fácil manejo y que responda eficientemente a las necesidades planteadas [10].

3. Objetivo General y Objetivos Específicos

3.1. Objetivo General

Desarrollar una metodología de visualización interactiva y eficaz de información en alta dimensión, usando un modelo de interacción y técnicas de reducción de dimensión, que presente un buen compromiso entre desempeño en la representación de los datos y costo computacional.

3.2. Objetivos Específicos

- Seleccionar técnicas representativas de reducción de dimensión, teniendo en cuenta criterios de desempeño y costo computacional, con el fin de determinar las más adecuadas para ser utilizadas en entornos de visualización de datos.
- Formular un modelo matemático de interacción usuario-computador que habilite al usuario para realizar la combinación y/o selección de métodos reducción de dimensión a través de la visualización interactiva de los datos.
- Diseñar una interfaz que permita al usuario realizar un análisis visual e interactivo de los datos representados en baja dimensión.

4. Grado de Innovación y Contribuciones

4.1. Contribuciones

A continuación se menciona las contribuciones que el desarrollo de esta tesis de Maestría aportaría a las áreas de visualización de información, reducción de dimensión, minería de datos, aprendizaje de máquina y reconocimiento de patrones:

- Seleccionar técnicas representativas de reducción de dimensión, teniendo en cuenta criterios de desempeño y costo computacional, con el fin de determinar las más adecuadas para ser utilizadas en entornos de visualización de datos.
- Formular un modelo matemático de interacción usuario-computador que habilite al usuario para realizar la combinación y/o selección de métodos reducción de dimensión a través de la visualización interactiva de los datos.
- Diseñar una interfaz que permita al usuario realizar un análisis visual e interactivo de los datos representados en baja dimensión.

5. Contexto

En este capitulo se presentan algunos conceptos preliminares de Big Data en la sección 5.1 con sus respectivas características, Tipos de datos, Las 6V del Big Data y Procesamiento en Datos Masivos y Visualización. En la sección 5.2.3 presenta las generalidades del campo de visualización, con sus propiedades y dominios. 5.3 se presenta los conceptos de minería de datos, el proceso de minería de datos y la visualización. La sección 5.4 discute el concepto global sobre la Reducción de la Dimensionalidad y su propósito. En la sección 5.5 se conceptualiza sobre la Visualización de los Datos y cómo esta ciencia encuentra una relación con las tareas de Reducción de la Dimensionalidad.

5.1. Concepto de Big Data

El tratamiento de grandes volúmenes de datos alcanza magnitudes exponenciales desde la primera década del siglo XXI. Esta tendencia permite dirigir la tecnología de los avances hacia nuevos paradigmas de la toma de decisiones, que se utiliza para entender los datos masivos (estructurados, no estructurados y semi-estructurados). Se evidencia un costo computacional muy alto, si se administra desde cualquier motor de base de datos relacional tradicional. La volatilidad de la información en el procesamiento y análisis de datos con herramientas tradicionales se llama Big Data [2] [3] [4].

5.1.1. Tipos de Datos y Fuentes de Big Data

La diversidad de datos generada puntualmente por fuentes como: personas, máquinas, transacciones, empresas entre otras. Se determina que dichas fuentes poseen la mayor abundancia en datos de organizaciones, cada una con diferentes áreas de producción que generan datos de tipo “no estructurados” [9]. Es por esto, que se consideran dichos datos, para ejercicios de análisis por parte de las empresas para enfocar esfuerzos en penetraciones de nuevos mercados, así como también en labores de toma de decisiones, sin dejar de lado la interconexión entre máquinas (M2M). Las transacciones y la biometría; generan grandes volúmenes de datos, tal como las señales de GPS que se generan, lecturas con RFID, datos emitidos por sensores, datos del sector salud, llamadas y demás registros; se trata de un crecimiento exponencial que día a día tiende a incrementarse [11]

[12]. A continuación se presenta en la tabla (**5-1**) la definición de los tipos de datos y un ejemplo de cada uno de ellos.

Tabla 5-1.: Tipos de Datos en Big Data

Tipos de Datos	Definición	Ejemplo
Estructurados	Datos con formato o esquema fijo que poseen campos fijos.	Hojas de cálculo y archivos o ficheros.
Semi-estructurados	Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores.	Textos de etiquetas XML y HTML.
No Estructurados	Datos sin tipos definidos, se almacenan principalmente como documentos u objetos sin estructura uniforme.	Audio, video, fotografia, formatos de texto libre (e-mails;SMS, artículos; libros; mensajería de tipo WhatsApp, Viber, etc.)

Joyanes, L. (2014). Big Data: análisis de grandes volúmenes de datos en organizaciones. Barcelona: Marcombo.

5.1.2. Las 6V del Big Data

Las organizaciones del mundo enfrentan mediante las herramientas de descubrimiento de conocimiento (KDD) los grandes volúmenes de datos, sin embargo no es el único desafío a enfrentar [13]. En particular, IBM y Gartner plantean el ”modelo de las tres V” (3V o V3) para referir las características propias de big data: volumen, velocidad y variedad. No obstante, a medida que esta tecnología y los volúmenes de datos crecen, se evidencia que; se sigue sumando nuevas características tales como la veracidad ”incluida por IBM como la cuarta característica así como otras fuentes añaden valor y visualización” [11].

Las 6V de Big Data, se ha convertido en un tema candente de nuestra realidad computacional, hoy día los Datos son lo suficientemente grandes para ser almacenados, administrados y analizados con las nuevas herramientas que brindan a los procesos comerciales la extracción de relevancia de información en tiempo real, debido a la gran velocidad de la circulación de los datos [14]. Una muestra de ello son, todos los ejercicios de econometría, el cual implementa muchos métodos cuantitativos y cualitativos para proyectar regiones de decisión en tiempo real bajo el marco de la nuevas tecnologías a nivel de desarrollo de software, ciencia de los datos y minería de datos entre otros [14] [15].

En tareas de representación de datos en matrices de alta dimensión, se estima que el proceso de tratamiento para comunicar los resultados debe ser la Visualización, el cual conlleva un gran esfuerzo académico en el proceso de interpretación de estos mismos, debido a la inter-disciplinariedad

de los conceptos a abordar [16]. No obstante con el equipo adecuado suele ser un poco más accesible. A continuación se hace una breve descripciónmediante mediante la tabla (**5-2**), acerca de las nuevas características de la Big Data [14] [17].

5.1.3. Big Data Procesamiento y Visualización

La gran capacidad de procesamiento de las máquinas y su bajo costo computacional de almacenamiento, permite que exponencialmente sea posible generar y agrupar datos masivamente, la información de estos grupos suele tergiversarse en muchas capas del cual la mayoría de casos es de gran importancia, pero con dificultad de acceder a ésta fácilmente [18] [19], El descubrimiento de conocimiento en bases de datos (DCBD o KDD) básicamente posee un proceso que a partir de éstos datos, combina el descubrimiento y análisis de información para conllevar a la extracción de patrones en forma de reglas o funciones, con la finalidad de garantizar al usuario la realización del análisis respectivo, DCBD es realizado a partir de las siguientes tres etapas: Pre procesamiento, (Data mining) y visualización de información. [18] [19] [20].

Comúnmente las técnicas de procesamiento de datos permiten recuperar información oculta en su totalidad, ya que dicha información es de gran importancia y necesaria para aplicar técnicas de recuperación de información como la minería de datos [18] [19]. En la segunda década del siglo XXI, el crecimiento de los datos ha generado una gran demanda en el desarrollo de procesos que permiten entender la información de estos volúmenes, esto se hace eficientemente mediante la minería de datos, pero los volúmenes de datos grandes pueden generar conjuntos o reglas similares. Estas formas de representación del conocimiento requieren de analistas con habilidades en la interpretación de patrones y extraer realmente el conocimiento inmerso [21]. Esta es una de las razones por las que surgen las técnicas de reducción de la dimensionalidad que permiten en cierta medida mitigar el problema de la dimensión para estos resultados, permitiendo reducir por ejemplo; 5000 variables a sólo 5 o 4, pero aun Así, tales variables pueden ser abstractas, así como también, estas técnicas necesitan un experto para su interpretación [22]. En la actualidad, se han desarrollado herramientas de visualización y visualización inteligentes que permiten comprender mejor el número de reglas grandes y los parámetros obtenidos de la aplicación de minería de datos, interactuando con múltiples presentaciones visuales con respecto a la información [23].

La información visual tiene un papel muy importante en la minería de datos, porque su objetivo es descubrir el conocimiento inmerso en los datos, tal conocimiento sólo puede ser determinado por los métodos de minería de datos, pero si el conocimiento es muy difícil de interpretar, aumenta el tiempo gastado, el dinero y la comprensión. (Una forma supone presenciar a un experto en materia). En general hay herramientas que incluyen etapas de pre-procesamiento, uso de métodos de minería de datos, post-procesamiento y / o visualización. Sin embargo muchas herramientas no integran todas las etapas, terminando en resultados abstractos de la información. Además, las

Tabla 5-2.: Las 6V del Big Data

Tipos de Datos	Definición	Ejemplo
Volumen	Las actuaciones diarias tanto de empresas como de personas usuarias generan grandes volúmenes de datos	Se hablaba de gigabytes, ahora se referencian petabytes y exabyte, para 2015 a 2020 será la era del zettabyte
Velocidad	Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores.	Flujos continuos de datos que son imposibles de manipular por sistemas tradicionales.
Variedad (tipos de datos)	Datos sin tipos definidos, se almacenan principalmente como documentos u objetos sin estructura uniforme.	Los datos de redes sociales, imágenes y videos pueden venir de sensores y no suelen estar preparados para una integración en una aplicación.
Veracidad	La veracidad o fiabilidad (truth) es la confianza y credibilidad que los datos generados por big data suponen en la toma de decisiones en las empresas.	A medida que la variedad y las fuentes de datos crecen la fiabilidad suele ser menor para los directivos de las organizaciones.
Valor	Las organizaciones estudian cómo obtener información de los grandes datos de una manera rentable y eficiente.	Tecnologías que faciliten la analítica de datos (las tecnologías de código abierto como Apache Hadoop), aportan valor a las organizaciones.
Visualizacion	Actualmente muchas de las imágenes que nos traen a la memoria el trabajo con big data tienen que ver con estas nuevas formas de ‘ver’ estos datos.	El exponencial crecimiento de la información genera cada vez más problemáticas en torno a la gestión de la privacidad de la información y la visualización de contenidos.

herramientas de integración para todas las etapas no tienen especial énfasis en la visualización, esta razón causa resultados ambiguos en el análisis visual [21] [24]. Mediante el uso de métodos de reducción de la dimensión puede transformar las representaciones visuales en objetos 1D, 2D o 3D que son más inteligibles para los seres humanos.

5.2. Concepto de INFOVIS

En la segunda mitad del siglo XVIII hasta hoy día, La visualización de la información, por sus siglas en ingles (INFOVIS), basa sus cimientos sobre dos principios fundamentales tales como: EL principio de reducción y El principio de las variables espaciales [25] [26], estos principios son la base de nuevos paradigmas de la representación visual como; la visualización directa que sobre escribe algunas reglas como el nuevo enfoque de permutación de primitivas por objetos reales en términos de visualización que aportan al proceso de análisis e investigación de patrones [26].

5.2.1. Concepto de Modismo (Idiom) en INFOVIS

Un modismo es una representación visual que posee una estructura de datos diseñada para presentar morfologías visuales, el cual tiene como característica fundamental el análisis y la validación del diseño visual de la tarea a resolver, en otras palabras un modismo es la representación visual (gráfico) o resultado visual de la tarea. Los modismos (Idiom), se pueden evaluar en nivel de expresividad según las magnitudes que tengan los canales del cual Tamara Munzner en su libro ”visualization analysis and design” del 2014 de University of British Columbia, pudo identificar los canales más expresivos teniendo en cuenta su efectividad en el ser humano bajo los modelos de cognición general, HCI y demás ciencias inmersas [27].

5.2.2. Principio de la Reducción y Variables Espaciales

El principio de la reducción, se ubica como el primero debido al uso de primitivas gráficas el cual dispone de líneas, rectas, puntas, formas y curvas geométricas simples para representar objetos y relaciones entre ellos [28]. El uso de variables espaciales tales como: (posición, tamaño, forma y movimiento). Se posiciona como el segundo principio fundamental, dado que dichas variables son especialmente útiles para establecer patrones y relaciones, así como también para representar diferencias claves en los datos [29].

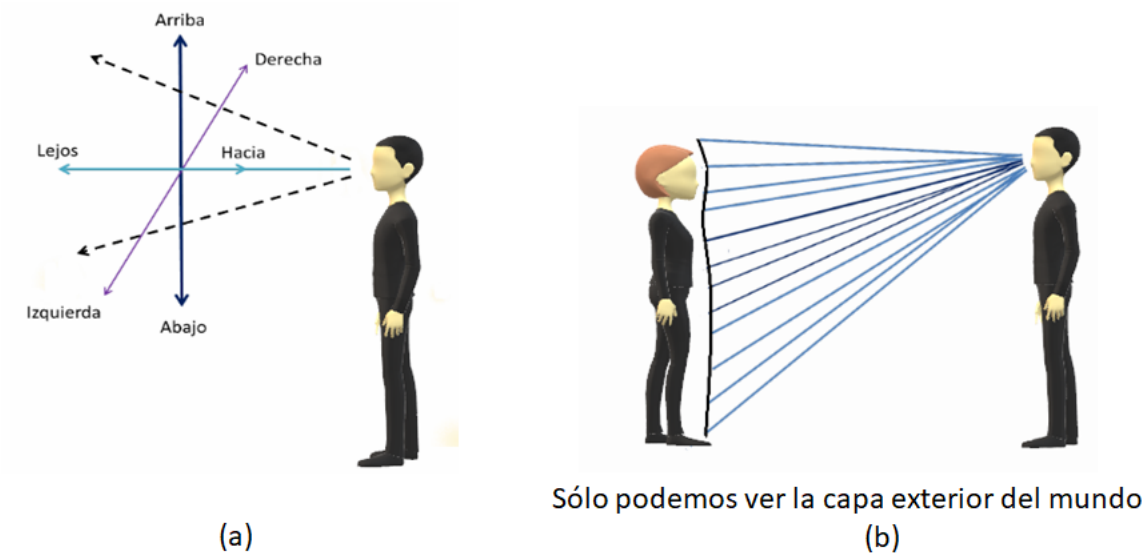


Figura 5-1.: Ancho de banda del campo visual humano. a) De lado y de arriba abajo los ejes son fundamentalmente diferentes del eje de profundidad. (b) A lo largo del eje de profundidad sólo podemos ver un punto por cada rayo, frente a millones de rayos para los otros dos ejes. Basado en [Tamara Munzner, with illustrations by Eamonn Maguire].

5.2.3. Concepto de Visualización y su Importancia

Las representaciones visuales de conjuntos de datos, proporcionadas por los sistemas de visualización basados en informática, se orientan en pro de ayudar al ser humano en la búsqueda de ejecutar tareas de una manera rápida y efectiva, utilizando la visión como el sentido más efectivo y con mayor ancho de banda (ver figura (5.2.3)) para aumentar las capacidades de percepción de nuestro entorno [27].

Los ejercicios de visualización son asertivos cuando se justifica la necesidad por parte del ser humano de aumentar sus capacidades sensoriales, en lugar de reemplazar al factor humano con modelos computacionales que ejercen la toma de decisiones. El enorme espacio de búsqueda de posibles diseños de visualización es incluyente, porque considera las metodologías para la creación e interactividad de las representaciones visuales. Por esta razón se establece que el diseño visual, al ser un universo de posibilidades se ve limitado solo por enmarcar representaciones dentro del espacio de diseño, esto implica que para algunas tareas en particular aquellas representaciones sean ineficientes. En virtud de ello la validación de la eficacia se hace necesaria [30] [31].

El diseñador de visualización debe concebir dentro de sus modelos, los 3 tipos de limitantes a nivel de recursos, como: “de pantallas”, “de hardware o computadoras” y finalmente “de seres humanos” [32]. El uso de la visualización, se analiza en referencia de la “necesidad del usuario”

comúnmente llamado el (¿porqué?), además de los datos a mostrar y como se diseña todas aquellas estructuras de datos o morfologías visuales de representación, a las que se bautizan como (Modismo) [33].

5.2.4. Dominios de la Visualización Versus la Automatización

La visualización permite que los humanos puedan realizar sencillos análisis de cualquier expresión generalizada de datos, especialmente cuando no saben puntualmente que preguntas necesitan de antemano [1]. En la segunda década del siglo XXI se promete una mejor toma de decisiones gracias a la era de los datos masivos, es por esta razón que la definición de; "tareas", "preguntas" o "problemas" sobre los datos, se espera en el mejor de los casos que sean preguntas bien definidas, dando espacio a modelos y técnicas computacionales de campos tales como; la estadística, la probabilidad y el aprendizaje automático [34] [35]. En la era moderna la automatización de procesos o tareas se hacen mediante una "solución basada en computación" comúnmente conocida como Computer-Based, esta clase de actividades anteriormente fueron desarrolladas por los seres humanos, pero hoy día, son soluciones totalmente automáticas que han sido aceptadas [36], puesto que proyectan la no dependencia del juicio humano y por tanto no hay necesidad de diseñar una herramienta visual, ya que el proceso es aceptable en términos de automatización y se sale de los parámetros del dominio de la identificación de necesidad de visualizar [37].

5.2.5. Apoyo al Aumento de las Capacidades Humanas como Herramienta Analítica

El 72% de problemas de análisis visual, están mal planteados o especificados: "las personas no saben cómo abordar el problema" [38]. Se parte del punto que existen docenas, millares o aun millones de posibles preguntas para preguntar, y las personas no tienen una certeza para determinar cuál de estas preguntas candidatas son correctas a priori o de antemano [39]. La mejor elección es impartir un proceso, aumentando las capacidades humanas para el ejercicio del análisis donde se pueda explotar las potentes propiedades de reconocimiento y detección de patrones del sistema visual humano [40]. Los sistemas visuales son asertivos para su uso y ejecución cuando su objetivo general es aumentar las capacidades humanas en términos sensoriales y demás, con el único propósito de no ser reemplazados completamente por un proceso automático carente de imprecisión debido al mal planteamiento o especificación inicial [41]. Las herramientas visuales pueden ser transitorias, con el ánimo de maximizar procesos tempranos de transición exploratorios, donde obtener una comprensión más clara de los requisitos de análisis juega un papel importante antes del desarrollo formal de modelos matemáticos, computacionales o soluciones automáticas [42] [43]. Esto permite que el resultado del diseño que concierne a las problemáticas concretas dentro del dominio real, suelen entenderse de una manera más clara articulando la resolución de las tareas del

usuario, así como también de la misma herramienta [44].

5.2.6. Análisis Exploratorio para el Descubrimiento Científico

Se pueden crear herramientas visuales el cual estén dirigidas a procesos exactos con el fin de interactuar netamente con modelos computacionales, el cual apoya fuertemente los procesos para refinar, depurar o ampliar los algoritmos de algún sistema o simplemente entender como los algoritmos se ven afectados por el ejercicio de cambios de parámetros [45]. El marco conceptual de este tipo de visualizaciones, está dirigido a otra audiencia que necesita un sistema superior que determine cuál de los múltiples modelos computacionales, matemáticos o algoritmos son más idóneos en determinadas circunstancias para que su salida pueda satisfacer el análisis exploratorio para el descubrimiento científico, cuyo objetivo es acelerar y mejorar la capacidad de un usuario para generar y comprobar hipótesis [46].

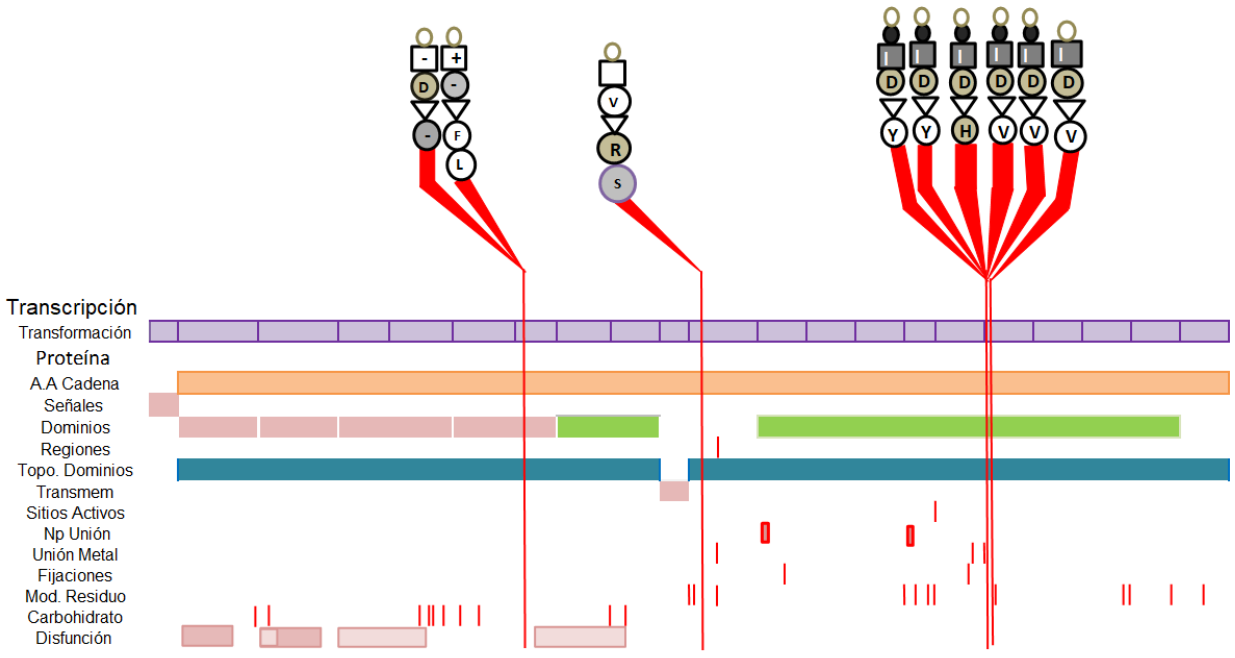


Figura 5-2.: Basado en la herramienta Visión de Variantes ayuda a los biólogos a evaluar el impacto de las variantes genéticas acelerando el proceso de análisis exploratorio. De [Fersta yet].

El caso más frecuente es el análisis exploratorio para el descubrimiento científico, cuyo objetivo es ”acelerar y mejorar la capacidad de un usuario para generar y comprobar hipótesis” [47]. La Figura (**5-2**) muestra una herramienta diseñada para ayudar a los biólogos a estudiar las bases genéticas

de la enfermedad a través del análisis de la variación de la secuencia de ADN [48]. Aunque estos científicos hacen un gran uso de la computación como parte de su flujo de trabajo más grande, no hay esperanza de automatizar completamente el proceso de la investigación del cáncer en el corto plazo [49].

5.3. Minería de Datos

Minería de datos (Data Mining): Se define como el proceso de descubrimiento de patrones, tendencias y relaciones significativas al examinar grandes volúmenes de datos para determinar la información sumergida (llamada información oculta) en dichos datos. Específicamente, las técnicas de minería de datos están dirigidas a patrones de descubrimiento, perfiles y tendencias de interés a través del análisis de datos mediante reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos. De acuerdo con los requerimientos de los usuarios, en la actualidad, la minería de datos ha permitido que estas técnicas y tecnologías penetren directamente en los entornos de base de datos actuales [19]. La minería de datos importa conceptos de Aprendizaje Automático, Inteligencia Artificial y estadísticas multivariantes para analizar los patrones en las bases de datos, donde éstas se representan en forma de arrays con información estructurada [20].

5.3.1. Proceso de Minería de Datos

Tal como se muestra en la imagen (**5-3**), la minería tradicional, el proceso de minería de datos se lleva a cabo en varias fases de desarrollo, el cual existen varias técnicas de extracción (en este caso, para extraer esa información considerada relevante para el negocio y mostrarla de un modo comprensible) [50].

Los procesos inclusivos en la minería de datos tradicional, establece como primer paso, la selección de los datos a tratar del cual posee una relación lineal con las variables de cálculo y predicción, en otras palabras, un conjunto de variables que actúan como función objetivo y otro de variables independientes de la siguiente manera: las variables función objetivo definen la elección en pro de sus objetivos que dependen del análisis, y las segundas determinan la manera mediante el cual el proceso se ejecutará [51] [52].

Acto seguido, se realiza un análisis de las propiedades de aquellos datos en selección para establecer el reconocimiento de tendencias, valores atípicos, patrones, así como también datos a descartar de tipo vacío o nulo que nada aportan al modelo [53]. De esta modo el procesamiento posterior de estos datos, permite impartir labores de clasificación y segmentación de trabajar articulada-mente con el modelo predictivo que se elige. Después se crean aquellos modelos de aprendizaje gracias

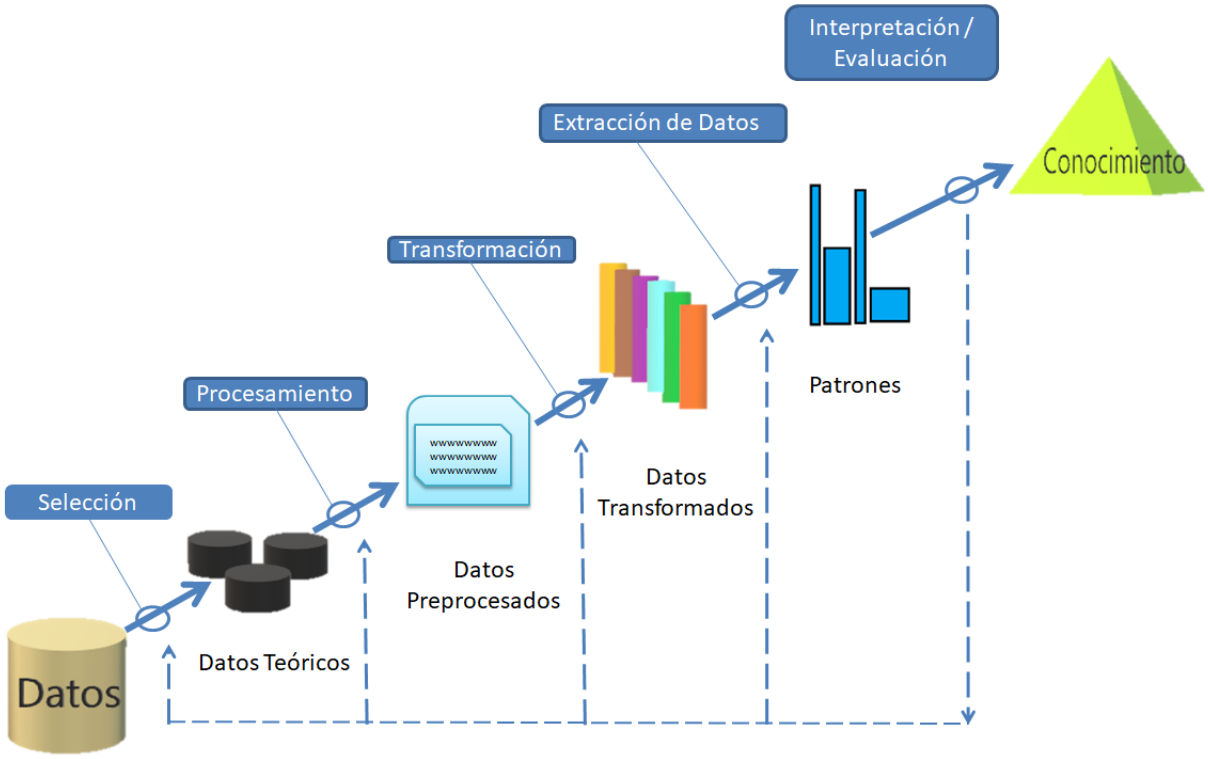


Figura 5-3.: Descripción del proceso general de minería de datos, en referencia a las 5 fases en desarrollo. Basado en [Han y Kamber, 2006].

al reconocimiento de patrones comporta-mentales, así como también las reglas de asociación, disociación que poseen las variables dentro del ámbito en desarrollo del análisis previo [54].

Finalmente cuando se obtienen los modelos de conocimiento aplicando dichas técnicas, se establece el siguiente paso que es la validación de estos, el cual ya se han comparado e interpretado para elegir el mejor según las métricas de resultados [55]. No obstante si en el evento que ningún modelo se ajusta a la resolución del problema o expectativas de lo esperado en términos de conocimiento, el proceso se repite de nuevo cambiando variables y adoptando técnicas distintas a las usadas en los procesos anteriores, de esta manera se obtiene un modelo donde se satisface las necesidades del conocimiento esperado [56].

5.3.2. La Visualización en Minería de Datos

En concreto para ahorrar tiempo, agilizar el proceso y esfuerzos a los expertos, que deben determinar en tiempo récord si los modelos obtenidos corresponden a lo esperado, se establece las herramientas que cumplen el rol de visualizar la información para cumplir con este cometido [57].

En los ejercicios de comparación de modelos así como las pruebas de evaluación mediante las métricas de calidad entre otras, para determinar el grado de satisfacción, es donde son protagonistas todas aquellas herramientas de visualización de datos [58] [59], que cumplen su papel de abstraer o simplificar de una manera ágil las tareas de aquellos expertos permitiendo que en tiempos muy cortos se optimicen los procesos de extracción de conocimiento, minimizando los riesgos asociados a los malos resultados debidos a las desacertadas interpretaciones que se puedan asociar [60]. Por consiguiente, la visualización de datos están íntimamente relacionados con una correcta gestión de los datos y la información [61] [62].

5.4. Reducción de la Dimensión

Un paso que se utiliza frecuentemente para el pre-procesamiento de datos es la reducción de dimensión (RD), que toma un subconjunto de variables para que el espacio de renderización original de los datos se reduzca óptimamente. Según ciertos criterios de calidad, cuyo objetivo será diferenciar el subconjunto que permite representar la mejor manera el espacio inicial. Los datos iniciales corresponden a muestras u objetos representados en características o variables. La inclusión de un gran número de variables dentro del proceso de exploración de datos puede aumentar los costos y el tiempo de procesamiento. Incluso puede generar datos con información redundante, ruidosa e irrelevante [8].

La reducción de dimensionalidad (RD) tiene como objetivo extraer información relevante en baja dimensión a partir de datos de alta dimensión, ya sea para mejorar el rendimiento computacional del sistema de reconocimiento de patrones o para permitir la visualización de datos de manera inteligible. Como aproximación a los métodos clásicos de RD, se encuentran el conocido análisis de componentes principales (PCA) y el escalamiento multidimensional clásico (CMDS), que se basan respectivamente en los criterios de varianza y la distancia de conservación [63].

Este trabajo se enmarca en un enfoque espectral. Las técnicas espectrales se han aplicado con éxito en varias tareas de reducción de dimensión, como el análisis de relevancia [64], [65] y la extracción de características [66], [67], entre otros.

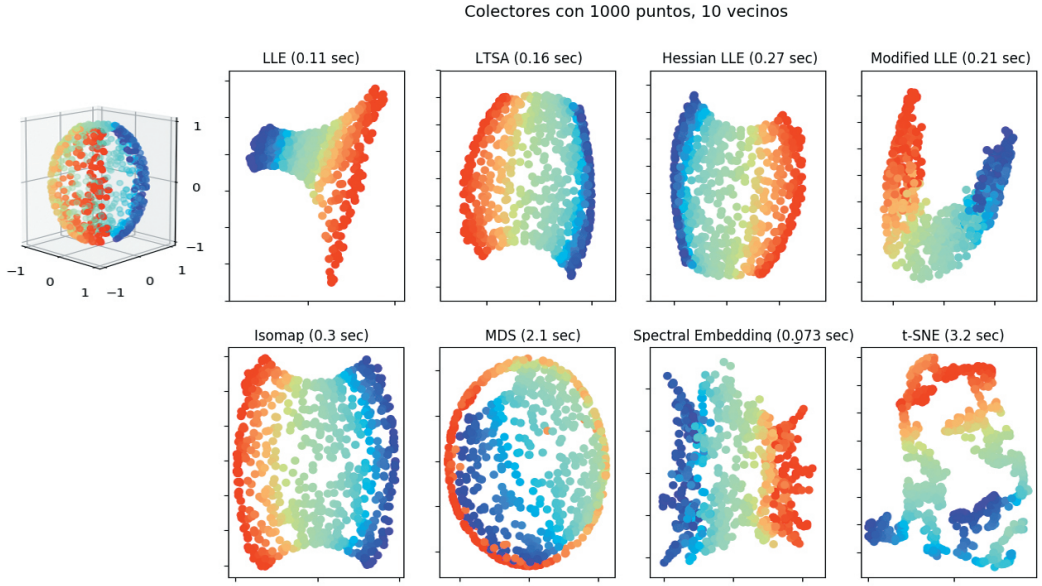


Figura 5-4.: Una ilustración de la reducción de la dimensión en *Estructura Esférica Artificial en 3D*, con varios métodos de reducción de la dimensión, generados mediante el Toolbox de reducción de la dimensión ManiFolds con Anaconda 3.6 - Python

En este trabajo se introducen enfoques de reducción de la dimensión espectral generalizada, se consideraron las implementaciones estándar de escalamiento multidimensional clásico (CMDS) [63], incrustación localmente lineal (LLE) [68], grafo Laplacian eigenmaps (LE) [69] entre otros. Además, sus aproximaciones de núcleos se consideraron [70]. El rendimiento de RD se cuantifica por una versión reducida de la aceptación media de K-barrios descritos en [71]. Como resultado, se ofrecen alternativas de métodos espectrales para la reducción de dimensiones tales como MDS, LLE y LE; Además de un marco versátil para explicar enfoques ponderados.

5.5. Visualización de Datos desde la Reducción de la Dimensión

Visualización de datos: La reducción de la dimensionalidad permite representar los datos originales en dimensiones inteligibles, en 2 o 3 dimensiones. Sin embargo, la percepción humana va más allá de una presentación especial y geométrica. Además de esto hay también otros factores importantes de estos aspectos y otros relacionados con el análisis visual de datos, como la visualización de datos (también, visualización de información - InfoVis). Esta área se ocupa de todo lo relacionado con la comunicación entre un usuario y la computadora, donde el factor más relevante es; la utilidad de la información procesada por ese ordenador en el futuro [72]. Dentro de la visualización de datos se utilizan interfaces interactivas cuyo principal objetivo es representar con mínima entropía visual una serie de datos al usuario final. La visualización de la información debe tener las siguientes características: Ser inter-relacional, transformar datos abstractos en información relevante, buscar la mínima pérdida de información en esta transformación e intuitivamente dirigirse a los usuarios que interactúan, transforman e interpretan esta información. Algunos estudios se han dedicado al desarrollo y al estudio de técnicas de visualización basadas en la forma geométrica [73], [74], coordenadas paralelas y radiales [75], [76], [77], [78], pixels [79], subspaces [80], iconos [23], entre otros enfoques [20], [81].

Como se mencionó anteriormente, una forma intuitiva de visualización de datos es, a través de gráficos 2D o 3D que resulta en una visualización natural e inteligible para los seres humanos. Por esta razón, la reducción de la dimensionalidad tiene relevancia, siendo una etapa importante tanto para la minería de datos, reconocimiento de patrones como para los sistemas de visualización de datos [82]. Técnicamente, la reducción de la dimensionalidad (RD) tiene como objetivo lograr la representación de datos en el espacio de baja dimensión, lo que mejora el desempeño de las tareas de minería de datos, aprendizaje de máquina, exploración de datos entre otras y las características de la representación general de datos, considerando su naturaleza intrínseca que son más adecuadas e inteligibles para el ser humano [83].

Parte II.

Estado del Arte

6. Estado del Arte de Métodos de Reducción de Dimensión

En este capitulo se presenta un estado del arte, considerando las vertientes de tipo espectral en la sección 6.1, acto seguido se presentan los métodos basados en divergencias en la sección 6.2, heurística en la sección 6.3. De esta manera se presentan las generalidades de los diferentes métodos asociados a cada categoría con sus respectivas características y componentes propios de funcionamiento. En la sección 6.4.3 se presenta una taxonomía de reducción de la dimension mediante un gráfico de jerarquías y sus correspondientes tablas (ver **6-1**, y **6-2** de clasificación de métodos de reducción de la dimensión, con los factores de clasificación de cada categoría. Finalmente se presenta la sección 6.5 de visualización usando reducción de la dimensión e interactividad.

6.1. Métodos Espectrales

6.1.1. Basados en Similitudes

Isomap

Posee una métrica geodésica, permitiendo medir distancias topológicamente basadas en grafos, el cual hace de éste, un método no lineal [84]. Se afirma como el más simple debido a implementar una optimización algebraica exacta, es decir, mediante simples operaciones algebraicas, se garantiza encontrar el óptimo global respecto a su función de error, además posee gran flexibilidad al no restringir proyecciones sobre un hiperplano [85] [86]. Isomap es un método espectral basado en similitudes debido a realizar la descomposición de una matriz de Gram en auto valores y vectores propios [87].

Locally Linear Embedding, LLE

Asume que los datos son linealmente locales, ya que esta propiedad permite proyectar los colectores de forma no lineal [84]. No obstante LLE, utiliza descomposición de valores y vectores propios siendo ésta una característica utilizada en modelos lineales. Sin embargo posee atributos

no lineales provenientes del cálculo de los K vecinos más cercanos permitiendo utilizar un procedimiento no lineal por perfección[88].

Laplacian Eigenmaps, LE

Puede verse como una variante de LLE, ya que sigue el mismo enfoque espectral calculando una matriz Gram que extrae los vectores propios que se asocian a los valores propios de menor valor [89] [90]. Sin embargo LE desarrolla un proceso local, con salvedad de abordar el problema no lineal de una manera diferente: LE se fundamenta en conceptos de teoría de grafos mediante la implementación de un operador laplaciano y el proceso de minimización de distancias locales (distancia de entre vecinos), donde el mapeo de estos puntos son llevadas a cabo a un solo punto (todas las distancias son cero)[91] [92] [93].

Hessian LLE (HLLE)

Modifica el modus operandi de Laplacian Eigenmaps, donde imparte un proceso de sustitución de la función cuadrática basada en el Laplaciano, por otra que se basa en una Hessiana. Aparte de esto, toma las técnicas de Locally linear embedding al utilizar una matriz dispersa, por tanto este método HLLE, deriva su conceptualización de isometría local, el cual proyecta un colector visto como un sub-colector de un espacio euclidiano, que permite recuperar los parámetros latentes de estos datos dispersos que se establecen sobre un colector embebido, en el espacio euclidiano de alta dimensión [94].

Diffusion Maps

Es conocido como un método espectral no lineal, debido a calcular las coordenadas a partir de valores y vectores propios aplicando un operador de difusión en los datos de entrada, este operador explota las relaciones entre la difusión y la aleatoriedad establecida por la cadena de markov [95]. En otras palabras la distancia euclidiana entre aquellos puntos del espacio embebido es congruente a la distancia de difusión entre las distribuciones de probabilidad enfocadas a estos puntos, por medio de la impartición de similitudes locales a diferentes escalas, que en virtud de ello, permite realizar una representación de baja dimensión global del conjunto de datos [96].

6.1.2. Métodos Basados en Disimilitudes

Principal Component Analysis, PCA

Generalmente este método determina el número de elementos descriptivos que subyace en un conjunto de datos que posee la mayor información de la varianza de estos [97] [98]. En otras palabras es equivalente decir que; busca la proyección de un conjunto de datos, el cual represente a estos de la mejor manera bajo el modelo de mínimos cuadrados, dicha proyección tiene un grado de correspondencia en términos de varianza acumulada por cada observación vista como un componente principal que mejor describe al conjunto de datos [99] [100] [101].

Multidimensional Scaling, CMDS

El principal propósito de este método radica en encontrar la topología que subyace en el conjunto de medidas de distancia entre los distintos objetos a mapear [102], por esta razón pretende utilizar un espacio de baja dimensión que establezca la mejor relación humana (2D o 3D), en términos de percepción y entendimiento [103] [104]. Estas observaciones se asignan específicamente a posiciones, de tal manera que aquellas distancias entre los puntos que están representadas en el espacio de baja dimensión sean congruentes en su máxima expresión con las disimilitudes iniciales [105] [106].

Linear Discriminant Analysis, LDA

Este método desarrolla una transformación lineal, que busca proyectar las muestras que siguen patrones de distribución gaussianas; lo anterior se establece con el fin de obtener resultados que comprometan una transformación embebida, que garantice la maximización de la varianza intra clase y extra clase [107] [108]. La tarea se define con el único propósito de realizar una robusta regla de decisión que pueda establecer un conjunto de datos a una baja dimensión mediante un vector que pueda garantizar el máximo en dispersión o separación entre las clases [109].

6.1.3. Basados en Kernel

Kernel PCA

En particular, PCA se fundamenta principalmente en las proyecciones lineales que preservan la mayor información, para representar de la mejor manera un elemento inmerso en el dominio de la varianza. Referente a la matriz de datos, en el caso de tener media cero respecto a las filas, dicha afirmación establece la preservación de varianza, el cual de otro modo, es equivalente a la preservación del producto interno euclidiano o producto escalar. Teniendo un espacio de representación

desconocido de alta dimensión tal que, este sea mayor que el original, en virtud de ello el cálculo del producto interno debe mejorar la representación y visualización de datos resultantes, por tanto se hace necesario una representación de tipo kernel, que calcule el producto interno euclidiano o producto escalar de este espacio desconocido de alta dimensión. El kernel básicamente es una función que mapea los datos de la dimensión original a otra mayor para mejorar notablemente las representaciones y visualización de aquellos datos resultantes [110].

Generalized Discriminant Analysis, KernelLDA

En solución a los diferentes problemas de LDA que concierne al concepto de linealidad, dicho método resulta ineficiente en espacios de alta dimensión [111]. En consecuencia de esto, surge KLDA que modifica este método tradicional al incluir aquellos procesos basados en Kernels, que transforman dramáticamente este método lineal en no lineal. Este nuevo enfoque, permite resolver el problema calculando la descomposición de valores propios. Sus aportes más significativos surgen en tareas de clasificación el cual logra una separabilidad en función de caracterizar los rasgos de dos o clases múltiples de objetos [112] [113].

6.1.4. Basados en Distancias

Análisis Factorial

Surge tras la necesidad de encontrar grupos homogéneos, llamados factores basados en la correlación lineal entre los elementos, con la finalidad de que estos grupos sean independientes de otros. Este método consta de las siguientes 4 etapas: El cálculo de la matriz que contiene la variabilidad de todos los atributos, el cálculo del número de factores óptimo, la rotación de la matriz para optimizar su interpretación y la estimación de las puntuaciones de los datos en el nuevo espacio de baja dimensión [114] [115].

6.2. Métodos Basados en Divergencias

6.2.1. Métodos Basados en Stochastic Neighbor Embedding (SNE)

SNE Convencional

Se establece como una aproximación probabilística que concede la representación a menor dimensión de objetos, que describen vectores en un espacio de alta dimensión o por disimilitudes proyectadas a un espacio de baja dimensión, que preserve las propiedades de vecindad, donde son evaluadas por la función de costo que utiliza la sumatoria de divergencias de Kullback-Leibler,

acto seguido se aplica un algoritmo de gradiente simple con la finalidad de ajustar las posiciones de la representación en el espacio de baja dimensión. Su modus operandi, consiste en asignar a cada objeto una probabilidad Gaussiana en el espacio original de alta dimensión, el cual; las densidades bajo esta probabilidad Gaussiana (o las disimilitudes dadas) son proyectadas para establecer distribuciones de probabilidad en todos los vecinos potenciales de un objeto. La finalidad del embebimiento es aproximar la distribución de probabilidad tanto como sea posible a la representación del objeto en baja dimensión [116].

t-Distributed SNE (tSNE)

Es un método no lineal que a diferencia de su antecesor SNE, que utiliza la metica de la distancia euclidiana como base de su similitud entre los puntos, tSNE lo realiza mediante una distribución de probabilidad. Este método particularmente posee un buen ajuste en el encrustamiento de datos de alta dimensión en un espacio de 2 o 3 dimensiones, para establecer su salida mediante un gráfico de dispersión. tSNE proyecta los puntos de datos similares en el espacio de alta dimensión para efectos de ser mapeados con la finalidad de medir aquellos puntos distantes. Concretamente se puede definir e dos pasos, teniendo como paso inicial la construcción de una distribución de probabilidad sobre el espacio original o de alta dimensión, puntualmente sobre aquellos puntos pares, por tanto aquellos puntos que comparten propiedades disimiles tienen muy poca probabilidad se ser seleccionados. Como paso final se establece una distribución de probabilidad similar sobre los puntos en el espacio de baja dimensión que reduce al mínimo la divergencia de Kullback-Leibler, respecto a las dos distribuciones de probabilidad [117] [118].

Symmetric Stochastic Neighbor Embedding (SymSNE)

Se generaliza a partir de la estructura de tSNE para implementar algunas mejoras sobre las funciones de encrustamiento por similitud. La idea principal surge sobre el uso de varias funciones de similitud que caracterizan funciones de puntuación negativa y que a partir de estas se pueda establecer un subconjunto de funciones parametrizadas de similitud, no obstante se hace necesario elegir la mejor función de similitud. Acto seguido se optimiza la función objetivo, el cual SymSNE realiza un proceso de optimización de punto fijo mediante el cual se aplica a todas las funciones y no necesita el factor humano para la configuración de algún parámetro. En estudios cuantitativos se afirma que, SymSNE es tan eficiente y rápido como tSNE, con la salvedad que agrega un plus re realizar una mejor separabilidad extra clase denominado (clusters) que su homologo tSNE [119].

6.2.2. Enfoque de Proximidad Embebido

Proximity Embedding (SPE)

Realiza las configuraciones iniciales de tal manera que proyecte iterativamente el ajuste de pares de objetos seleccionados al azar, del mismo modo refine las coordenadas de tal manera que las distancias en el proceso de mapeo estén ligadas con aquellas proximidades [120]. Las magnitudes de dichos refinamientos son controladas por un atributo o parámetro que evalúa la tasa de aprendizaje, que a su vez va disminuyendo en el lapso de la ejecución del algoritmo con finalidad de evitar el descontrol mediante un comportamiento oscilatorio. EN otras palabras SPE es robusto, sencillo y divergente [121].

6.3. Métodos Heurísticos

6.3.1. Redes Neuronales Artificiales

Self-Organizing Maps, SOM

Reconocido en la literatura por ser un algoritmo claro y relativamente rápido, de clase bio-inspirada en las famosas redes neuronales artificiales de mapa auto organizado de características que se puede entrenar mediante el aprendizaje no supervisado con la finalidad de producir la representación discreta del espacio de las muestras de entrada, tal como lo hace el perceptrón multicapa (MLP) [122]. Matemáticamente se expresa mediante una función de error que en apariencia resulta bastante intuitiva y fácil de entender en términos del procedimiento que dicho método realiza. Las propiedades de este algoritmo, permiten la combinación de dos procesos tales como: la representación topográfica vista como reducción de dimensión y la cuantización vectorial [123]. En términos de ingeniería se puede describir el proceso como; tomar el conjunto de datos, el cual se ajusta a un hiperplano codificándolo como un sistema de coordenadas de dicho hiperplano, teniendo en cuenta que las nuevas formas que puedan proyectar cambios en su morfología o estructura de datos, dependen linealmente si la condición de la dependencia lineal se cumple, en otras palabras, si una de ellas esta curvada la otra podrá hacerlo del mismo modo [124] [125].

Análisis de Componentes Curvilíneos (CCA)

Emerge como el pionero de los métodos en realizar una combinación de la reducción de la dimensión no lineal con el proceso de cuantización vectorial que ha sido lograda por la preservación de la distancia [126]. Este algoritmo comparte estructuras similares en su proceso, tal cual lo hace Self-Organizing Maps (SOM), el cual se inspira en argumentos biológicos utilizando redes neuronales artificiales (RNA) [127]. Por otra parte define el criterio de encrustamiento en función de preservar las distancias que posteriormente incluye la cuantización vectorial, además de proyectar

procesos de aproximación optimizada [128]. En conjuntos de datos pequeños así como también en colectores dispersos, este método resulta ser ineficiente debido a no aprovechar en plenitud la información disponible, por este motivo se recomienda que en tareas de conjuntos en menor proporción no sea utilizada la cuantización vectorial [129].

Curvilinear Distance Analysis, CDA

Se desarrolla como un método que generaliza el modelo de Análisis de Componentes Curvilíneos (CCA), cuyo atributo diferencial es la métrica geodésica, que establece el cálculo de las distancias a través de topologías basadas en grafos. Se considera como un prototipo de red neuronal de arquitectura artificial RNA, teniendo en cuenta que la cuantización vectorial es equivalente a las neuronas. Sin embargo la sustitución de la distancia euclidiana por otra basada en grafos, se puede asimilar como todas aquellas adiciones sinápticas en términos de conexiones neuronales, donde la estructura de conexión posee una gran similitud a una red de tipo (lattice) es decir, se garantiza primero el grafo como estructura de datos que morfológicamente corresponde a implementaciones de forma rectangular o hexagonal, haciendo limitada su aplicación debido a que pocos colectores pueden ajustarse a estas formas tan elementales, que debido a esto la estructura de la malla pueden resultar deformada en su totalidad al tratar de adaptarse a su conjunto de datos [130].

Isotop

Es inevitable la comparación entre Isotop y SOM, debido a la relación que tienen; tanto en su proceso, que se puede interpretar como una marcha atrás de self organizing maps SOM, como en la cuantización vectorial, que para este método se establece como un parámetro opcional. Ambos métodos basan su modus operandi, como modelos no lineales que usan la cuantización vectorial, además pertenecen al grupo de los algoritmos bio-inspirados tales como las redes neuronales artificiales (RNA), que además utilizan técnicas de optimización basadas en aproximaciones. Este algoritmo emerge de la necesidad de mejorar sustancialmente las limitaciones de SOM que puntualmente enfoca sus esfuerzos en el proceso de la reducción de la dimensión no lineal. Isotop se describe en proceso mediante 3 pasos como: Establecer la forma embebida en un espacio de baja dimensión (Cuadricula bidimensional de puntos regularmente espaciados), como paso siguiente la construcción de un lattice entre dos puntos y finalmente la cuantización vectorial en un espacio de alta dimensión [131] [132].

6.3.2. Deep Learning

Deep Autoencoders

Es una red neuronal profunda que se compone por dos redes simétricas, el cual se denominan redes de creencia profunda (Bilief Networks), son máquinas restringidas de Boltzmann (RBM), que son bloques de construcción de redes de creencia profunda, del cual cada capa RBM, básicamente no se comunican lateralmente con los nodos de una sola capa. Un ejemplo de ello es tener cierta cantidad de entradas del cual la primera capa del autoencoder posee ligeramente más parámetros, esto puede parecer un arma de doble filo, por la premisa de tener más parámetros que entradas, sin embargo es una buena manera de ajustar una red neuronal. Para este caso de estudio la maximización de estos parámetros o la ampliación de estas características de las entradas, hace posible proyectar la descodificación final de los datos auto-codificados. Visto de otro modo, el ancho de las capas va reduciendo los parámetros en función de las transformaciones de las unidades de creencia, esto indica que se implementara un proceso de partición por unidades de mitad hasta llegar al optimo global donde se produce un vector denominado auto codificado profundo, para distribuirse en la mitad para pre entrenamiento y el excedente como producto de un RBM normal, en lugar de una capa de salida en función de una clasificación mediante una función logística, como normalmente lo hace al final del proceso una red de creencia profunda [133].

6.3.3. Redes Neuronales Bayesianas

Mapas Topográficos Generativos (GTM)

Considerada como principal alternativa de self organizing maps SOM, se define como una red de densidad especifica basada en un modelo generativo, donde todas las variables del problema están latentes en una distribución de probabilidad a partir de un proceso bayesiano, comúnmente denominado red bayesiana [134]. Este tipo de redes proceden de manera diferentes dentro del esquema de aprendizaje, debido a que el aprendizaje tradicional tal como SOM lo realiza, asume una distribución de probabilidad sobre la probabilidad condicional del espacio de búsqueda, para determinar los valores óptimos que generalmente se encuentran con el máximo estimador de probabilidad [135]. Desde otra perspectiva el aprendizaje bayesiano en redes de densidad como GTM, toma una distribución de probabilidad que resulta del modelo inmerso en los parámetros de los datos, antes de considerar cualquier otro [136]. En otras palabras, instancia la probabilidad marginal de un dato que permite realizar una posible consideración sobre el valor del mismo, para después actualizar la anterior distribución a una nueva utilizando el teorema de bayes [137].

6.4. Otros Métodos

6.4.1. Sammon's Non Linear Mapping, NLM

Se creó con el fin de establecer un mapeo entre el espacio de alta dimensión y el espacio de baja dimensión, el cual ha sido denominado mapeo no lineal de Sammon (Sammon's no lineal mapping, NLM) [138]. Sin embargo este mapeo no es puntualmente un mapeo continuo entre los dos espacios cartesianos, porque su propósito y función principal es la reducción de la dimensión de un conjunto determinado o finito de datos [139]. Se dice que NLM comparte algunas funciones con MDS tales como; el embebimiento que realiza, así como también ninguno de estos métodos asume algún tipo de modelo genérico como lattice o estructuras de datos orientadas a grafos [140].

6.4.2. NLM Geodésico (GNLM)

Se asume como una generalización del NLM, con la diferencia que las métricas utilizadas para las distancias como; la distancia euclidiana, se cambia por la métrica de distancias geodésicas [141]. Esto es posible gracias al principio de la siguiente afirmación: la asignación no lineal de Sammon (NLM), se usa con un recurso muy importante con la distancia euclidiana, tanto en el espacio de datos como en el espacio embebido, no obstante, se asume que nada de lo anterior se establece como una restricción para definir otra métrica que corresponda al espacio de datos [142].

6.4.3. Taxonomía de Reducción de la Dimensión

A continuación se presenta como alternativa de entendimiento, un gráfico **6-1**, con la taxonomía de los diferentes métodos de Reducción de la Dimensión, así como también un resumen modo de tabla (ver Tablas **6-1**, y **6-2**.

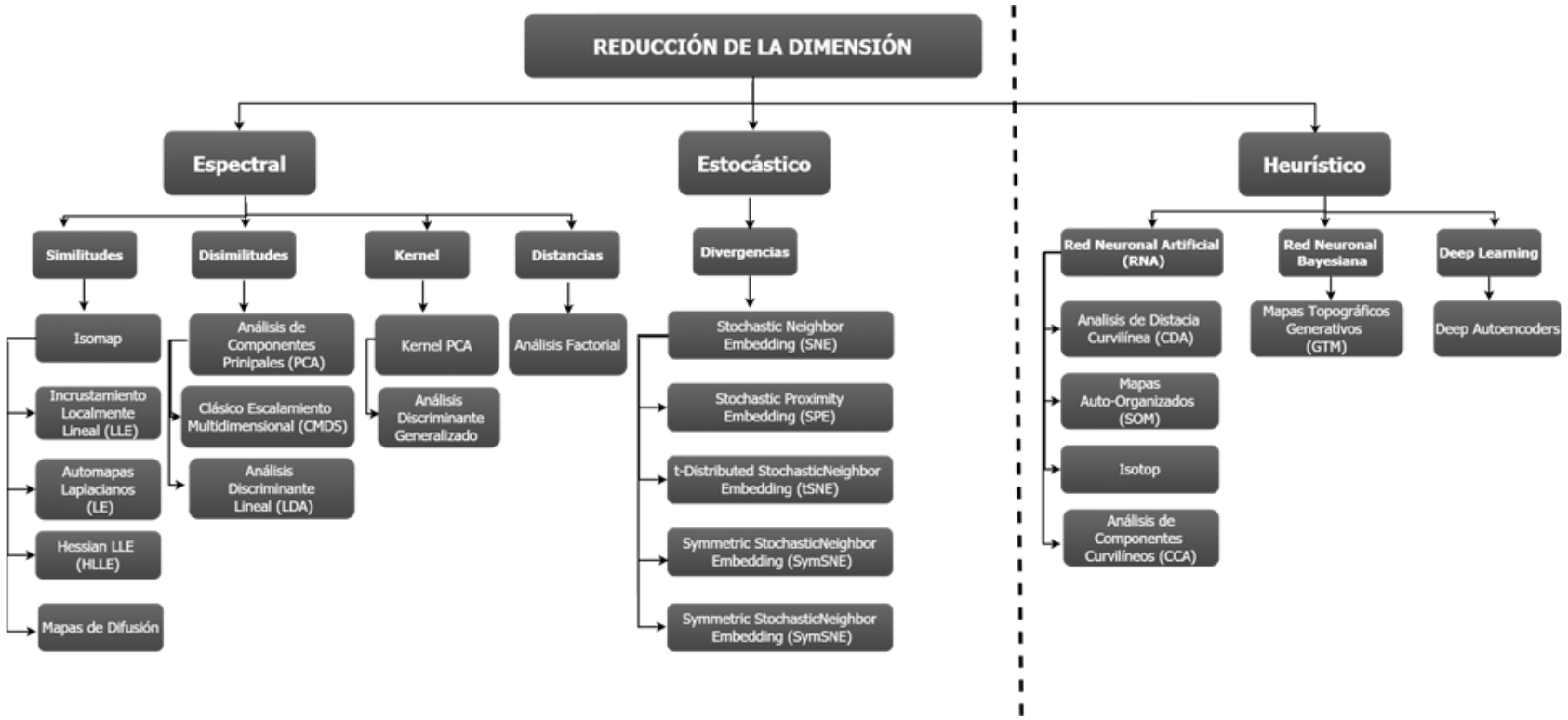


Figura 6-1.: Esta taxonomía, muestra la categorización de los métodos del paradigma no-lineal, más utilizados dentro del estado del arte actual (2017).

Tabla 6-1.: Clasificación de Métodos de Reducción de Dimensión de Tipo Espectral

Reduccion de la Dimensionalidad	Referencia Bibliografica	Clasificación	Sub-clasificación
Isomap	[84], [85], [86], [87]	Espectral	Basado en similitudes
Incrustamiento Localmente Lineal (LLE)	[84], [88]	Espectral	
Automapas Laplacianos (LE)	[89], [90], [91], [92], [93]	Espectral	
Hessian LLE (HLLE)	[94]	Espectral (Isometrico)	
Mapas de Difusión	[95], [96]	Espectral	
Analisis de Componentes Prinipales (PCA)	[97], [98], [99], [100], [101]	Espectral (Lineal, estadístico)	Basado en disimilutdes
Clasico Escalamiento Multidimensional (CMDS)	[102], [103], [104], [105], [106]	Espectral (Lineal)	
Analisis Discriminante Lineal (LDA)	[107], [108], [109]	Espectral	
Kernel PCA	[110]	Espectral	Kernel
Analisis Discriminante Generalizado	[111], [112], [113]	Espectral	
Analisis Factorial	[114], [115]	Espectral (Lineal, estadístico)	Basado en distancias

Tabla 6-2.: Clasificación de Métodos de Reducción de Dimensión Basados en Divergencias, Heurísticos y otros

Reduccion de la Dimensionalidad	Referencia Bibliografica	Clasificación	Sub-clasificación
Stochastic Neighbor Embedding (SNE)	[116]	Estocástico	Basado en divergencias
Stochastic Proximity Embedding (SPE)	[120], [121]	Estocástico	
t-Distributed Stochastic Neighbor Embedding (tSNE)	[117], [118]	Estocástico	
Symmetric Stochastic Neighbor Embedding (SymSNE)	[119]	Estocástico	
Mapeo No Lineal de Sammon (NLM)	[138], [139], [140]	Mapeos no lineales	Otros
NLM Geodésico (GNLM)	[141], [142]	Mapeo no lineal (distancia geodésica)	
Analisis de Distacia Curvilinea (CDA)	[130]	Heurístico	Red Neuronal Artificial (RNA)
Mapas Auto-Organizados (SOM)	[122], [123], [124], [125]	Heurístico	
Isotop	[131], [132]	Heurístico	
Mapas Topograficos Generativos (GTM)	[134], [135], [136], [137]	Heurístico	Red Neuronal Bayesiana
Deep Autoencoders	[133]	Heurística	Deep Learning
Analisis de Componentes Curvilineos (CCA)	[126], [127], [128], [129]	Heurístico (espectral)	Red Neuronal Artificial (RNA)

6.5. Visualización Usando Reducción de Dimensión e Interactividad

La visualización de datos multidimensionales posee como elemento fundamental, las técnicas de reducción de la dimensión (RD). Se afirma que los métodos asociados a RD, pueden ser útiles bajo ciertas condiciones especiales. Estos características se asumen como restricciones debido a proyectar; el tratamiento de datos a nivel de análisis exploratorio, el cual espera surtir cambios dramáticos de adaptación a las necesidades humanas, así como también a problemas del contexto, idealmente en términos de interactividad y procesos sobre la marcha.

La mayoría de sistemas de análisis visual del estado del arte actual, demuestran los beneficios de integrar la Visualización Interactiva con la Reducción de la Dimensión. No obstante se hace necesario la comprensión general y estructura de esta integración. Los resultados de esta investigación revelan 3 escenarios comunes, que subyacen de procesos de interacción para concebir susceptibilidad en ejercicios de control interactivo como: restricciones algorítmicas, selección de características y la selección de la idoneidad entre una gama de algoritmos de RD [143].

6.5.1. Enfoque de Restricciones Algorítmicas

En el primer escenario se enmarca las restricciones algorítmicas, donde se estudian las afirmaciones de autores que defienden la tesis de; (La interacción visual con la reducción de la dimensionalidad en complejidades de tiempos de ejecución) [144]. El desarrollo de tal afirmación, evidencia los esfuerzos de centrar las investigaciones en implementaciones ingenieriles específicas de sistemas de análisis visual integrando RD, de manera que resultan en el análisis de las formas, que otros métodos de aprendizaje de máquinas han sido combinados con RD en pro de los costos computacionales, puestos en marcha desde una perspectiva interactiva enfocada netamente en la visualización [144].

Otros autores [145] [144], defienden conceptos que comparten algunas características pero con leves discrepancias. A la hora de proyectar el mejor ajuste en términos de percepción para la visualización, tal ajuste satisface aquellas representaciones de baja dimensión en beneficio de la exploración de la separabilidad de clases, así como también la distribución espacial de los datos que carecen del criterio que identifica las capacidades perceptivas de los humanos. Este concepto hace muy complejo e ineficaz para estructuras de clases complejas. Debido a esto, se enfocan a impulsar la percepción para maximizar las capacidades de observación, netamente desde la visualización, en las proyecciones inmersas en separación de clases, teniendo en cuenta el cálculo de modelos de cuantización vectorial y geométricos que optimizan el costo computacional [145].

6.5.2. Enfoque en Selección de Características

En este segundo escenario se evidencia una fuerte inclinación de algunos autores [143] [146], a la hora de ejecutar los métodos de RD en beneficio de un conocimiento previo. Con la única finalidad de establecer la esperanza de agrupamiento natural y selección de atributos, el cual logra optimizar esfuerzos humanos en la investigación de la idoneidad del factor del pre-procesamiento de los datos. Para lograrlo, identifican técnicas de agrupamiento natural que sean menos costosas en términos de computación y puedan llegar a feliz término, con un panorama general de analítica visual que pueda dar un sentido de orientación, tanto en la percepción visual de los seres humanos como salvaguardar el equilibrio del poder en la computación [146].

6.5.3. Enfoque de Idoneidad en Selección del Algoritmo

Algunos autores [147] [148], desarrollan especial esfuerzo en todos los movimientos intermedios, específicamente en la interpolación de los movimientos en las formas de la visualización. Se espera que los algoritmos de RD no convexos, bajo cambios de métricas del espacio de entrada puedan tener una correspondencia, en su proceso de interactividad que permita manejar los parámetros en función de un factor humano, es decir el usuario mediante una interfaz adecuada. Este enfoque tiene la bondad de concentrar esfuerzos en la relación intrínseca de los grupos de variables seleccionados dinámicamente, así como la evaluación del impacto en una sola variable o agrupamientos de variables en la topología de los datos [147].

Otros autores como [149] [148], proponen los algoritmos de RD en función de nuevas fases de exploración, en términos de configuraciones paramétricas. Esto se desarrolla con el único propósito de realizar proyecciones más asertivas del análisis visual de datos dimensionales. Dicha labor de sintonización de parámetros permite a estos usuarios una comparación cuantitativa en tiempo en marcha, el cual resulta ser más comprensible debido a mostrar con detalles, los movimientos internos de cada transformación realizada por las distintas proyecciones propias de cada método, para facilitar el proceso de selección del método más idóneo [149].

Parte III.

Marco Teórico

7. Marco Teórico

Esta sección presenta las generalidades del proceso de Reducción de la Dimensión en la sección 7.1, acto seguido, en esta sección se organizan las temáticas de la siguiente manera: En la sección 7.2 se realiza una descripción del background de los componentes que permiten calculo matricial de los métodos espectrales y métodos basados en mezclas de divergencias [150, 151], el cual se utilizaron para efectos de esta investigación, referente al proceso de los métodos ($Singular - Value - Decomposition$ - SVD Sección 7.2.1), en la sección 7.2.2 ($Eigenvalue - Decomposition$ - EVD) y en la sección 7.2.3 ($SquareRoot - Of - a - SquareMatrix$ - SRSM). Se hace una revisión de los métodos que componen el calculo matricial de dichos modelos, así como también aquellos factores de embebimiento propios de cada método.

Las secciones 7.3 y 7.4 esbozan los métodos estudiados. Por tanto, la mayoría de los nuevos enfoques espectrales que pueden ser fácilmente comprendidos dentro de la teoría de grafos. A menudo, tal grafo se construye como un grafo no dirigido y ponderado, del cual los puntos de datos representan los nodos, y una matriz de similitud simétrica no negativa (también afinidad) representa el peso de pares entre nodos. Esta formulación también es útil para determinar los clusters de datos subrayados dentro de los datos de entrada. [152].

7.1. Generalidades de Reducción de la Dimensión

La reducción de la dimensión (DR) permite extraer información relevante y de menor dimensión de grandes colecciones de datos con el fin de mejorar el rendimiento de un sistema de reconocimiento de patrones o permitir una visualización de datos inteligible. En otras palabras, el objetivo de la reducción de la dimensión es representar una matriz de datos de alta dimensión $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1 \leq i \leq N}$, tal que $\boldsymbol{y}_i \in \mathbb{R}^D$, en una matriz de dimensión menor $\boldsymbol{X} = [\boldsymbol{x}_i]_{1 \leq i \leq N}$, con $\boldsymbol{x}_i \in \mathbb{R}^d$, donde $d < D$. Los enfoques clásicos de la RD fueron concebidos siguiendo un criterio intuitivo, como la preservación de la varianza (análisis de componentes principales - PCA) o la preservación a distancia (escalamiento multidimensional clásico - CMDS) [153]. En la segunda década del siglo XXI, los métodos más desarrollados y recientes tienen como objetivo preservar la topología de los datos. Esta topología es a menudo dada por un grafo relacionado con los datos, construido como un grafo no dirigido y ponderado, en el que los puntos de datos representan los nodos, y una matriz de similitud no negativa (también denominada afinidad), que contiene los pesos de las aristas en ambos sentidos. Esta representación se explota tanto con métodos espectrales como basados en

divergencias. Por un lado, para los enfoques espectrales, la matriz de similitud puede representar el factor de ponderación para distancias pares, como sucede en Laplacian eigenmaps [154]. Por otra parte, una vez normalizado, también puede representar una distribución de probabilidad. Este último es el caso de los métodos basados en divergencias tales como stochastic neighbour embedding [155].

7.2. Métodos para el Cálculo de Matrices

7.2.1. Descomposición de Valores Singulares

La descomposición del valor singular (*SingularValueDecomposition* - SVD) de un M-por-N matriz $\boldsymbol{A}$ se escribe como:

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T, \tag{7-1}$$

donde $\boldsymbol{V}$ es una matriz ortonormal (o unitaria) de M-por-M tal que $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{M\times M}$, $\boldsymbol{\Sigma}$ es una matriz pseudodiagonal con el mismo tamaño que $\boldsymbol{A}$; las M entradas de σ_m en la diagonal se llaman los valores singulares de $\boldsymbol{A}$ y $\boldsymbol{U}$ es una matriz ortonormal (o unitaria) N-por-N de tal manera que $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{N\times N}$.

El número de valores singulares diferentes de cero, denota el rango de $\boldsymbol{A}$. Cuando el rango es igual para P (Dimensionalidad del espacio latente que suele ser $\mathbb{R}^P$), el SVD puede ser usado para calcular el (pseudo) inverso de $\boldsymbol{A}$:

$$\boldsymbol{A}^+ = \boldsymbol{U}\boldsymbol{\Sigma}^+\boldsymbol{V}^T, \tag{7-2}$$

donde la (pseudo) inversa de $\boldsymbol{\Sigma}$ se calcula trivialmente al transponerlo e invirtiendo sus entradas diagonales σ^m. La SVD se utiliza en muchos otros contextos y aplicaciones. Por ejemplo, el análisis de los componentes principales [97] [98], puede llevarse a cabo utilizando un SVD. Por cierto, cabe destacar que PCA utiliza un SVD ligeramente moderado. Asumiendo que $M < N$, $\boldsymbol{U}$ podría llegar a ser grande cuando $M \ll N$, que a menudo ocurre en PCA, y $\boldsymbol{\Sigma}$ contiene muchos ceros inútiles. Esto motiva una desconexión alternativa de la SVD, llamada SVD de tamaño económico, donde sólo se calculan las primeras columnas P de $\boldsymbol{U}$. Consecuentemente, $\boldsymbol{U}^T$ tiene el mismo tamaño que $\boldsymbol{A}$ y $\boldsymbol{\Sigma}$ se convierte en una matriz diagonal cuadrada. Una desconexión similar está disponible cuando $M > N$.
Para una matriz cuadrada y simétrica, la SVD es equivalente a la descomposición del valor propio (EVD; ver a continuación).

7.2.2. Descomposición de Valores Propios

La descomposición del valor propio (*EigenvalueDecomposition* - EVD) de una matriz cuadrada $\boldsymbol{A}$ de M-por-M, se escribe como

$$\boldsymbol{AV} = \boldsymbol{V\Lambda} \tag{7-3}$$

- Donde $\boldsymbol{V}$ es una matriz cuadrada M-por-M cuyas columnas $\boldsymbol{v}_m$ son vectores de norma unitaria llamados vectores propios de $\boldsymbol{A}$.
- $\boldsymbol{\Lambda}$ es una matriz diagonal M-por-M que contiene los valores propios M λ_m de $\boldsymbol{A}$.

La EVD es a veces llamada la descomposición espectral de $\boldsymbol{A}$. La ecuación (7-3) traduce el hecho de que los propios vectores mantienen su dirección después de la multiplicación izquierda por$\boldsymbol{A}$: $\boldsymbol{A}_{\boldsymbol{v}_m} = \lambda_m \boldsymbol{v}_m$. Además, el factor de escala es igual al valor propio asociado. El número de valores propios diferentes de cero da el rango de $\boldsymbol{A}$, y el producto de los valores propios es igual al determinante de $\boldsymbol{A}$. Por otra parte, la traza de $\boldsymbol{A}$, denotado tr ($\boldsymbol{A}$) y desviado como la suma de sus entradas diagonales, es igual a la suma de sus propios valores

$$tr(\boldsymbol{A}) \triangleq \sum_{m=1}^{M} a_{m,m} = \sum_{m=1}^{M} \lambda_m \tag{7-4}$$

En el caso general, aunque $\boldsymbol{A}$ contenga sólo entradas reales, $\boldsymbol{V}$ y $\boldsymbol{\Lambda}$ pueden ser complejas. Si $\boldsymbol{A}$ es simétrica ($\boldsymbol{A} = \boldsymbol{A}^T$), entonces $\boldsymbol{V}$ es ortonormal (los vectores propios son ortogonales además de estar normados); el EVD se puede volver a escribir como

$$\boldsymbol{A} = \boldsymbol{V\Lambda V}^T, \tag{7-5}$$

y los valores propios son números reales. Además, si $\boldsymbol{A}$ es positivo definido, entonces todos los valores propios son positivos. Si $\boldsymbol{A}$ es semidefinida positiva, entonces todos los valores propios no son negativos. Por lo tanto, una matriz de covarianza es semidefinida positiva.

7.2.3. Raíz Cuadrada de una Matriz Cuadrada

La raíz cuadrada de una matriz diagonal se calcula fácilmente aplicando la raíz cuadrada únicamente en las entradas diagonales. En comparación, la raíz cuadrada de una matriz cuadrada no diagonal puede parecer más difícil de calcular. En primer lugar, hay dos maneras diferentes de separar la raíz cuadrada de una matriz cuadrada $\boldsymbol{A}$. La primera definición asume que

$$\boldsymbol{A} \triangleq (\boldsymbol{A}^{\frac{1}{2}})^T(\boldsymbol{A}^{\frac{1}{2}}) \tag{7-6}$$

Si $\boldsymbol{A}$ es simétrico, entonces la descomposición del valor propio (EVD) de la matriz ayuda a volver al caso diagonal. La descomposición del valor propio (ver sección 7.2.2) de cualquier matriz simétrica $\boldsymbol{A}$ es

$$\boldsymbol{A} = \boldsymbol{V\Lambda V}^T = (\boldsymbol{V\Lambda}^{\frac{1}{2}})(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T) = (\boldsymbol{A}^{\frac{1}{2}})^T(\boldsymbol{A}^{\frac{1}{2}}), \tag{7-7}$$

donde $\boldsymbol{\Lambda}$ es diagonal. Si $\boldsymbol{A}$ es también positivo definido, entonces todos los valores propios son positivos y las entradas diagonales de $\boldsymbol{\Lambda}^{\frac{1}{2}}$ siguen siendo números reales positivos. (Si $\boldsymbol{A}$ es sólo semidefinida positiva, entonces la raíz cuadrada ya no es única). La segunda y más general definición de la raíz cuadrada se escribe como

$$\boldsymbol{A} \triangleq (\boldsymbol{A}^{\frac{1}{2}})(\boldsymbol{A}^{\frac{1}{2}}) \tag{7-8}$$

Una vez más, la descomposición del propio valor conduce a la solución. La raíz cuadrada se escribe como

$$\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^{-1}, \tag{7-9}$$

y es fácil comprobar que

$$\begin{aligned} \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}} &= \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^{-1} \\ &= \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^{-1} \\ &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} = \boldsymbol{A} \end{aligned} \tag{7-10}$$

Esto es válido en el caso general, es decir, $\boldsymbol{A}$ puede ser complejo y/o asimétrico, produciendo valores propios y vectores propios complejos. Si $\boldsymbol{A}$ es simétrica, la última ecuación puede simplificarse aún más, ya que los vectores propios son reales y ortonormales ($\boldsymbol{V}^{-1} = \boldsymbol{V}^{T}$). No es digno de mención que la segunda definición de la raíz cuadrada de la matriz pueda ser generalizada para calcular poderes matriciales:

$$\boldsymbol{A}^{p} = \boldsymbol{V}\boldsymbol{\Lambda}^{p}\boldsymbol{V}^{-1} \tag{7-11}$$

7.3. Métodos Espectrales

7.3.1. Classical Multidimensional Scalling (CMDS)

Para entender CMDS se analiza los componentes semánticos en su expresión más básica. Como acercamiento intuitivo, se conceptualiza el termino *"Escalamiento"* el cual, hace referencia a todo aquel proceso que se establece en un espacio métrico objetivo, para configurar los puntos a partir de la información que subyace en las distancias de interpunto. En segundo lugar, el termino que establece la tarea *"Multidimensional"*, realiza el proceso de escalamiento, sí y solo sí el espacio objetivo es euclidiano [156]. El enfoque clásico de CMD, denominado por sus siglas *"CMDS - métrico clásico"*, se afirma que no es un enfoque de conservación de la distancia debido a garantizar los productos escalares en lugar de las distancias [157]. El MDS métrico clásico dentro de su proceso es incapaz de lograr la reducción de dimensión de forma no lineal. Sin embargo, el MDS métrico es un método del cual basa sus cimientos en la estructura de un modelo generativo simple [158]. En términos del método analítico su ejecución establece una permutación o cambio de eje ortogonal que separa las variables observables o triviales en $\mathbf{y}$ del mismo modo aplicado en variables latentes, almacenadas en $\mathbf{x}$:

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{7-12}$$

del cual los componentes de $\mathbf{x}$ son independientes y $\mathbf{W}$ es un D-por-P matriz tal que $\mathbf{W}^T\mathbf{W} = \mathbf{I}_P$. Para aquellas variables observadas como aquellas latentes. Aplicadas a un conjunto finito de puntos N, escritos en forma de matriz como

$$\mathbf{Y} = [..., \mathbf{y}(i), ..., \mathbf{y}(j), ...] \quad , \tag{7-13}$$

visto de otro modo, se puede proponer una reducida notación tal que, para el producto escalar entre vectores $\mathbf{y}(i)$ y $\mathbf{y}(j)$:

$$s_{\mathbf{y}}(i, j) = s(\mathbf{y}(i), \mathbf{y}(j)) = \langle \mathbf{y}(i) \cdot \mathbf{y}(j) \rangle \quad , \tag{7-14}$$

tal cual, se ha hecho para las distancias se puede escribir que

$$\begin{aligned} \mathbf{S} &= [s_{\mathbf{y}}(i, j)]_{1 \leq i,j \leq N} = \mathbf{Y}^T\mathbf{Y} \\ &= (\mathbf{W}\mathbf{X})^T(\mathbf{W}\mathbf{X}) \\ &= \mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{X} \\ &= \mathbf{X}^T\mathbf{X}. \end{aligned} \tag{7-15}$$

Donde $\mathbf{Y}$ y $\mathbf{X}$ son desconocidos; sólo la denominada matriz de Gram de productos escalares a pares $\mathbf{S}$, se da. Para el cálculo de los valores en las variables latentes se puede realizar trivialmente calculando la descomposición del valor propio (ver sección 7.2.2), de la matriz Gram $\mathbf{S}$: Al desconocer tanto $\mathbf{Y}$ como $\mathbf{X}$; sólo se define la matriz de productos escalares pares $\mathbf{S}$, llamada matriz de Gram. Como se mencionó anteriormente se calcula la descomposición de los valores propios (ver sección

7.2.2), en pro de hallar los valores de las variables latentes, el cual resulta ser una solución trivial para la matriz de Gram **S**:

$$\begin{aligned}\mathbf{S} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \\ &= (\mathbf{U}\mathbf{\Lambda}^{1/2})(\mathbf{\Lambda}^{1/2}\mathbf{U}^T) \\ &= (\mathbf{\Lambda}^{1/2}\mathbf{U}^T)^T(\mathbf{\Lambda}^{1/2}\mathbf{U}^T) \quad ,\end{aligned} \tag{7-16}$$

donde **U** es un *N*-por-*N* matriz ortonormal y $\mathbf{\Lambda}$ es un *N*-por-*N* diagonal que posee los valores propios. (La razón por la que **U** se utiliza en lugar de **V** como en la sección (7.2.2), se debe a la comparación que se realizará con su homólogo PCA en la siguiente sección. Vale la pena resaltar que **S** es la matriz Gram de los datos centrados, los valores en *D*hacen referenciaa la dimensionalidad del espacio de los datos $\mathbb{R}^D$, donde los valores propios son estrictamente positivos mientras que otros son cero en $\mathbf{\Lambda}$.) Si se clasifican en orden descendente los valores propios, la estimación del valor de *P*-Las variables de latencia dimensional, se calculan como un producto de la siguiente manera

$$\hat{\mathbf{X}} = \mathbf{I}_{P\times N}\mathbf{\Lambda}^{1/2}\mathbf{U}^T. \tag{7-17}$$

Equivalencias entre CMDS y PCA

En muchos trabajos del estado del arte actual [159], se menciona casi obligatoriamente sobre las equivalencias de MDS métrico y PCA. En esta tesis de maestría, se parte de la solución de la ecuación (7-17) para proyectar las equivalencias entre MDS métrico y PCA que se demostraran a continuación. Se afirma que, el MDS métrico y el PCA ofrecen la misma solución. Para efectos de ésta demostración, se toman las coordenadas de los datos en **Y**, que se asumen como conocidas. Esto es una regla in-negociable en el caso del método de *Análisis de Componentes Principales - PCA*, no obstante en el caso de MDS métrico, se observa que es lo opuesto y centrado. Además, la descomposición del valor singular (7.2.1) de **Y** se puede escribir como $\mathbf{Y} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T$. Entonces se puede afirmar que PCA descompone la matriz de covarianza, que es proporcional a $\mathbf{Y}\mathbf{Y}^T$, en vectores y valores propios:

$$\hat{\mathbf{C}}_{yy} \propto \mathbf{Y}\mathbf{Y}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}_{PCA}\mathbf{V}^T \quad , \tag{7-18}$$

donde la división por *N* se omite intencionalmente en la covarianza, y $\mathbf{\Lambda}_{PCA} = \mathbf{\Sigma}\mathbf{\Sigma}^T$. La solución es $\hat{\mathbf{X}}_{PCA} = \mathbf{I}_{P\times D}\mathbf{V}^T\mathbf{Y}$(referencia libro). Por otro lado, el MDS métrico descompone la matriz Gram en eigenvectores y valores propios:

$$\mathbf{S} = \mathbf{Y}^T\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}_{MDS}\mathbf{U}^T \quad , \tag{7-19}$$

donde $\mathbf{\Lambda}_{MDS} = \mathbf{\Sigma}^T\mathbf{\Sigma}$. La solución es $\hat{\mathbf{X}}_{MDS} = \mathbf{I}_{P\times D}\mathbf{\Lambda}_{MDS}^{1/2}\mathbf{U}^T$. Al igualar ambas soluciones y al volver a utilizar la descomposición de los valores singulares de **Y**:

$$\begin{aligned}\hat{\mathbf{X}}_{PCA} &= \hat{\mathbf{X}}_{MDS} \\ \mathbf{I}_{P\times D}\mathbf{V}^T\mathbf{Y} &= \mathbf{I}_{P\times D}\mathbf{\Lambda}_{MDS}^{1/2}\mathbf{U}^T \\ \boldsymbol{I}_{P\times D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T &= \boldsymbol{I}_{P\times D}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})^{\frac{1}{2}}\boldsymbol{U}^T \\ \boldsymbol{I}_{P\times D}\boldsymbol{\Sigma}\boldsymbol{U}^T &= \boldsymbol{I}_{P\times D}\boldsymbol{\Sigma}\boldsymbol{U}^T\end{aligned} \tag{7-20}$$

Esto demuestra que PCA y MDS métrico llegan a minimizar el mismo criterio.
De esta manera, para el caso del MDS métrico, puede ser reescrito como:

$$E_{MDS} = \sum_{i,j=1}^{N} (s_y(i,j) - \langle \hat{\mathbf{x}}(i) \cdot \hat{\mathbf{x}}(j) \rangle)^2 \quad . \tag{7-21}$$

La equivalencia entre los dos métodos puede ser una ventaja en algunas situaciones.

- Cuando los datos consisten en distancias o similitudes, la ausencia de las coordenadas no nos impide aplicar el PCA: basta con sustituir el PCA por el MDS métrico.
- Cuando se conocen las coordenadas, la equivalencia es también muy útil cuando el tamaño de la matriz de datos es muy grande, por ende $\mathbf{Y}$ se vuelve problemático.
- Si los datos no son demasiado dimensionales pero el número de puntos es enorme, PCA gasta menos recursos de memoria que MDS ya que el producto. $\mathbf{Y}\mathbf{Y}^T$ tiene un tamaño más pequeño que $\mathbf{Y}^T\mathbf{Y}$.

Respecto al proceso del algoritmo CMDS, el cálculo de las distancias en par requiere $O(N^2)$ entradas de memoria y $O(N^2D)$ operaciones. En realidad, las complejidades temporales y espaciales del MDS métrico están directamente relacionadas con las de un EVD. Computar todos los valores y vectores propios de una matriz no espesa de N-por-N típicamente exige lo siguiente $O(N^3)$ operaciones, dependiendo de la implementación.

Algoritmo (Classical Multidimensional Scalling - CMDS).

---

1. Si los datos disponibles consisten en vectores recogidos en $\mathbf{Y}$, y luego centrarlos, calcular los productos escalares en ambos sentidos. $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$, y vaya al paso 3.
2. Si los datos disponibles consisten en distancias euclidianas a ambos lados, transfórmelos. en productos escalares:
   - Cuadrado de las distancias y construir $\mathbf{D}$.
   - Realizar el doble centrado de $\mathbf{D}$, esto produce $\mathbf{S}$.
3. Calcular la descomposición del valor propio $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.
4. Una representación P-dimensional se obtiene computando el producto $\hat{\mathbf{X}} = \mathbf{I}_{P\times N}\mathbf{\Lambda}^{1/2}\mathbf{U}^T$.

Redes Basadas en Datos - (Data-Driven Lattice)

Los métodos que se describen en la siguiente sección 7.3.2 y 7.3.3 en oposición a los métodos que utilizan una red predefinida, no asumen la forma y la topología de la representación dimensional menor. Por el contrario, define la información contenida en los datos para establecer la topología del conjunto de datos y calcular la forma del embebimiento. Por tanto, dicho embebimiento no se limita de ninguna manera y puede adaptarse para capturar la forma del colector. En los métodos LLE y LE, detallados en las próximas secciones, la red de datos se formaliza mediante un grafo cuyos vértices son los puntos de datos y cuyos bordes representan las aristas o relaciones vecinales [159].

7.3.2. Locally Linear Embedding (LLE)

LLE propone un enfoque basado en transformaciones conformes. Dicho en otras palabras, posee cualidades gráficas que se podrían interpretar como un "malla o rejilla rectangular" o "mapa conformal", tal que preserva los movimientos ángulos locales [160]. La preservación de los ángulos locales, puede interpretarse como una manera de preservar los productos escalares locales, del mismo modo las distancias locales realizan un trabajo similar que está estrictamente relacionado, pero resulta ser otra manera diferente de preservar los escalares locales [161] [162].

Este método, dentro de su proceso inicia la construcción de un mapa conformal, el cual determina los ángulos primarios a tener en cuenta, el cual selecciona un par de vecinos en función de sus puntos tales como: $\mathbf{y}(i)$ en el data set $Y = [..., \mathbf{y}(i), ..., \mathbf{y}(j), ...]_{1 \leq i,j \leq N}$. Si se conoce la estructura del colector subyacente, entonces se puede asumir que un valor de K existe de tal manera que el colector es aproximadamente lineal, en la escala local de los *K-ary* vecindarios. La idea de LLE es reemplazar cada punto $\mathbf{y}(i)$) con una combinación lineal de sus vecinos. El error de reconstrucción total se puede medir con la función de coste cuadrático simple:

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{N} \left\| \mathbf{y}(i) - \sum_{j \in \mathcal{N}(i)} w_{i,j} \mathbf{y}(j) \right\|^2 \quad , \tag{7-22}$$

donde $\mathcal{N}(i)$ contiene a todos los vecinos del punto $\mathbf{y}(i)$ y $w_{i,j}$ las entradas de la matriz N-por-N $\mathbf{W}$, a los vecinos en la reconstrucción de $\mathbf{y}(i)$. Brevemente, $\varepsilon(\mathbf{W})$ suma todas las distancias cuadradas entre un punto y su reconstrucción local lineal. Para calcular los coeficientes $w_{i,j}$, la función de costes se reduce al mínimo con dos limitaciones. Los puntos son reconstruidos únicamente por sus vecinos, es decir, los coeficientes. $w_{i,j}$ y para puntos fuera del vecindario de $\mathbf{y}(i)$ son iguales a cero: $w_{i,j} = 0 \forall j \notin \mathcal{N}(i)$. Las filas de la matriz del coeficiente suman a una $\sum_{j=1}^{N} w_{i,j} = 1$. LLE busca que estas propiedades geométricas principalmente puedan ser válidas para representaciones de baja dimensión de los datos.

Puntualmente, LLE supone que los datos se encuentran cerca de un colector liso, no lineal de baja dimensionalidad. Y luego, a una buena aproximación, que consiste en una traslación, rotación y escala, del cual mapea las coordenadas de alta dimensión de cada vecindario a las coordenadas globales intrínsecas. Los pesos de reconstrucción $w_{i,j}$ reflejan las propiedades geométricas intrínsecas de los datos que son invariables a exactamente estas transformaciones. En el último paso de LLE, cada punto de datos de alta dimensión se mapea a un vector de baja dimensión que representa las coordenadas globales intrínsecas en el colector. Esto se hace seleccionando las coordenadas P-dimensionales para minimizar la función de coste de la representación de baja dimensión:

$$\Phi(\hat{\mathbf{X}}) = \sum_{i=1}^{N} \left\| \hat{\mathbf{x}}(i) - \sum_{j \epsilon \mathcal{N}(i)} w_{i,j} \hat{\mathbf{x}}(j) \right\|^2 \quad , \tag{7-23}$$

Esta función de coste, el cual conservar propiedades similares a la Ec. (7-22), suma los errores de reconstrucción causados por la reconstrucción local lineal. En este caso, los errores se calculan en el espacio de representación de baja dimensión y los coeficientes. $w_{i,j}$ son fijos. La minimización de $\Phi(\hat{\mathbf{X}})$ da las coordenadas de baja dimensión $\hat{\mathbf{X}} = [..., \hat{\mathbf{x}}(i), ..., \hat{\mathbf{x}}(j), ...]_{1 \leq i,j \leq N}$ que la mejor reconstrucción $\mathbf{y}(i)$ otorgada $\mathbf{W}$. En la práctica, la minimización de las dos funciones de costes $\varepsilon(\mathbf{W})$ y $\Phi(\hat{\mathbf{X}})$ se realiza de la siguiente manera: Primero, pueden ser calculados en forma cerrada los coeficientes restringidos de $w_{i,j}$, para cada punto de datos por separado. Considerando un punto de datos en particular $\mathbf{y}(i)$ con K vecinos más cercanos, su contribución a $\varepsilon(\mathbf{W})$ es:

$$\varepsilon_i(\mathbf{W}) = \left\| \mathbf{y}(i) - \sum_{j \in \mathcal{N}(i)} w_{i,j} \mathbf{y}(j) \right\|^2 \quad , \tag{7-24}$$

que puede ser reformulado como

$$\varepsilon_i(\omega(i)) = \|\mathbf{y}(i) - \sum_{r=1}^{K} w_r(i) \nu(r)\|^2 \tag{7-25}$$

$$= \| \sum_{j=1}^{K} \omega_r(i)(\mathbf{y}(i) - \nu(r))\|^2 \tag{7-26}$$

$$= \sum_{r,s=1}^{K} \omega_r(i)\omega_s(i) g_{r,s}(i) \quad , \tag{7-27}$$

donde $\omega(i)$ es un vector que contiene las entradas no cero de la fila i-ésimo (sparse) de $\mathbf{W}$ y $\nu(r)$ la r-ésimo vecino de $\mathbf{Y}(i)$, correspondiente a $\mathbf{y}(j)$ en la notación de Eq. (7-24). La segunda igualdad se mantiene gracias a la restricción (reformulada) $\sum_{r=1}^{K} \omega_r(i) = 1$, y el tercero usa el K-por-K matriz de Gram local $\mathbf{G}(i)$ cuyas entradas se definen como

$$g_{r,s}(i) = (\mathbf{y}i - \nu(r))^T (\mathbf{y}i - \nu(s)) \quad . \tag{7-28}$$

Las matrices $\mathbf{G}(i)$ pueden ser interpretadas como una especie de matrices de covarianza local alrededor de $\mathbf{y}(i)$. Utilizando un multiplicador Lagrange puede minimizarse en forma cerrada el error de reconstrucción, para reforzar la restricción $\sum_{r=1}^{K} \omega_r(i) = 1$. La inversa de $\mathbf{G}(i)$, los pesos óptimos son dados por

$$\omega_r(i) = \frac{\sum_{s=1}^{K}(\mathbf{G}^{-1}(i))_{r,s}}{\sum_{r,s=1}^{K}(\mathbf{G}^{-1}(i))_{r,s}} \quad . \tag{7-29}$$

Esta solución requiere una inversa de la matriz de covarianza local. Otra manera de minimizar el error, es simplemente resolver el sistema lineal de ecuaciones. $\sum_{r=1}^{K} g_{r,s}\omega_r(i)$ y luego escalar los coeficientes para que se sumen a uno, dando el mismo resultado. Por construcción, la matriz $\mathbf{G}(i)$ es simétrica y positiva semidefinida. Desafortunadamente, puede ser singular o casi singular, por ejemplo; cuando hay más vecinos que las dimensiones en el espacio de datos. ($K > D$). En este caso, $\mathbf{G}$ puede condicionarse, antes de resolver el sistema, añadiendo un pequeño múltiplo de la matriz de identidad:

$$\mathbf{G} \leftarrow \mathbf{G} + \frac{\triangle^2 tr(\mathbf{G})}{K}\mathbf{I} \quad , \tag{7-30}$$

donde $\triangle$ es menor en comparación con el rastro de C. Esto equivale a penalizar las grandes ponderaciones que explotan las correlaciones más allá de cierto nivel de precisión en el proceso de muestreo de datos. En realidad, $\triangle$ es de alguna manera un parámetro "oculto" de LLE. La minimización de la segunda función de costes $\Phi(\hat{\mathbf{X}})$ en pro de resolver un problema propio. Para este propósito, $\Phi(\hat{\mathbf{X}})$ se desarrolla de la siguiente manera:

$$\Phi(\hat{\mathbf{X}}) = \sum_{i=1}^{N} \left\| \hat{\mathbf{x}}(i) - \sum_{j \epsilon \mathcal{N}(i)} w_{i,j}\hat{\mathbf{x}}(j) \right\|^2 \tag{7-31}$$

$$= \sum_{i=1}^{N} \left\| \sum_{j \epsilon \mathcal{N}(i)} w_{i,j}(\hat{\mathbf{x}}(i) - \hat{\mathbf{x}}(j)) \right\|^2 \tag{7-32}$$

$$= \sum_{i,j=1}^{N} m_{i,j}(\hat{\mathbf{x}}(i)^T \hat{\mathbf{x}}(j)) \quad , \tag{7-33}$$

donde $m_{i,j}$ es la entrada de un N-por-N matriz $\mathbf{M}$, definido como

$$\mathbf{M} = (\mathbf{I} - \mathbf{w})^T(\mathbf{I} - \mathbf{W}) \quad , \tag{7-34}$$

que es simétrico y positivo semidefinido. La optimización se realiza en base a las restricciones que hacen que el problema esté bien planteado. Las coordenadas $\hat{\mathbf{x}}(i)$ se pueden traducir como un desplazamiento constante sin afectar al coste. Este grado de libertad desaparece si las coordenadas

están centradas en el origen. ($\sum_{i=1}^{N} \hat{\mathbf{x}}(i) = 0$). Además, a fin de evitar soluciones degeneradas, las coordenadas latentes se limitan a tener covarianza unitaria. ($\hat{\mathbf{C}}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T = \mathbf{I}$). Tal restricción simplemente explota la invariancia de la función de coste a las rotaciones y a las recalificaciones homogéneas. La integración óptima, hasta una rotación global del espacio de inserción, se obtiene calculando el fondo $P + 1$ vectores propios de la matriz $\mathbf{M}$. El último vector propio de $\mathbf{M}$, que LLE descarta, es un vector de unidad escalado con todos los componentes iguales; representa un modo de conversión libre y está asociado con un valor propio cero. Descartar este eigenvector refuerza la restricción de que las representaciones de menor dimensión tienen una media cero, ya que los componentes de otros eigenvectores deben sumar cero en virtud de la ortogonalidad con el último. El resto P los vectores propios dan el estimado P coordenadas dimensionales de los puntos $\hat{\mathbf{x}}(i)$ en el espacio latente.

Algoritmo (Locally Linear Embedding - LLE).

---

1. Para cada dato $\mathbf{y}(i)$,calcular
   - la K vecinos más cercanos de $\mathbf{y}(i)$,
   - la matriz regularizada $\mathbf{G}(i)$ según la Ec. (7-28) y (7-30),
   - los pesos $\omega(i)$ (Ec. (7-29)).
2. Conociendo los vectores $\omega(i)$, construir las matrices dispersas $\mathbf{W}$ y $\mathbf{M}$ (Ec. (7-34)).
3. Calcule el EVD de $\mathbf{M}$; las coordenadas estimadas están dadas por los vectores propios asociados con el segundo para (1+P)-ésimo valores propios más pequeños.

7.3.3. Laplacian Eigenmaps (LE)

Este método de reducción de la dimensión pertenece a la familia no lineal, actualmente muy desarrollado, basando sus cimientos en el ejercicio de la descomposición espectral [163]. El método tenía por objeto remediar algunas deficiencias de otros métodos espectrales como LLE, descritos en la sección (7.3.2) [164].

LE comparte similitudes dentro de su modus operando, como establecer una cercana relación con LLE, no obstante aborda el problema de una manera diferente, por ejemplo: en lugar de reproducir pequeñas correcciones lineales alrededor de cada dato, LE se basa en conceptos grafo-teóricos como el operador laplaciano en un grafo [165]. LE se basa en la minimización de distancias locales, es decir, distancias entre puntos de datos vecinos. Para evitar la solución trivial donde todos los puntos son mapeados a un solo punto (todas las distancias son entonces cero!), la minimización

es limitada. LE se basa en una hipótesis única y sencilla: el conjunto de datos:

$$\mathbf{Y} = [..., \mathbf{y}(i), ..., \mathbf{y}(j), ...]_{1 \leq i,j \leq N}, \tag{7-35}$$

posee un gran número de puntos N cerca de una superficie lisa denominada P-el colector. Como sólo se indican los puntos de datos, el colector en sí mismo permanece desconocido. Sin embargo, si N tiene un gran tamaño, el múltiple subyacente puede ser representado con buena precisión por un gráfico $G = (V_N,E)$. En esta representación, un vértice v_i del gráfico se asocia con cada punto de referencia $\mathbf{y}(i)$, y una arista conecta los vértices v_i y v_j si los puntos de datos correspondientes son vecinos, hablando en términos de grafos. Aquellas relaciones de vecindad pueden determinarse utilizando K-ary vecindarios o ϵ-ball como para otros métodos basados en grafos. Las relaciones vecinales pueden codificarse en una estructura de datos específica o simplemente en una matriz de adyacencia $\mathbf{A}$. Las entradas binarias $a_{i,j} \in \{0, 1\}$ indican si los puntos de datos $\mathbf{y}(i)$ y $\mathbf{y}(j)$ son vecinos o no. El objetivo de LE es mapear $\mathbf{Y}$ a un conjunto de puntos de baja dimensión:

$$\mathbf{X} = [..., \mathbf{x}(i), ..., \mathbf{x}(\mathbf{j}), ...]_{1 \leq i,j \leq N}, \tag{7-36}$$

que mantienen las mismas relaciones de vecindad. A tal concepto, se le entenderá por criterio definido:

$$E_{LE} = \frac{1}{2} \sum_{i,j=1}^{N} ||\mathbf{x}(i) - \mathbf{x}(j)||_2^2 w_{i,j}, \tag{7-37}$$

donde las entradas $w_{i,j}$ de la matriz simétrica $\mathbf{W}$ se relacionan con los de la matriz de adyacencia de la siguiente manera: $w_{i,j} = 0$ sí $a_{i,j} = 0$; de lo contrario, $w_{i,j} \geq 0$. Existen varias opciones posibles para las entradas distintas de cero. Algunos autores como; Mikhail Belkin y Partha Niyogi [referencia], recomiendan para estos casos utilizar un Kernel en forma de campana gaussiana:

$$w_{i,j} = exp\left(-\frac{||\mathbf{y}(i) - \mathbf{y}(j)||_2^2}{2T^2}\right), \tag{7-38}$$

donde el parámetro T puede considerarse como una temperatura en un kernel involucrado en ecuaciones de difusión. Una opción más simple consiste en tomar $w_{i,j} = 1$ sí $a_{i,j} = 1$. Esto equivale a fijar $T = \infty$ en el kernel. Según la definición de $\mathbf{W}$, minimizando E_{LE} bajo las limitaciones apropiadas es un intento de garantizar que si $\mathbf{y}(i)$ y $\mathbf{y}(j)$ están cerca el uno del otro, entonces $\mathbf{x}(i)$ y $\mathbf{x}(j)$ debería estar cerca también. En otras palabras, se conservan las propiedades topológicas (es decir, las relaciones de vecindad) y los pesos. $w_{i,j}$ actúan como sanciones que son más pesadas para puntos de datos cercanos. De acuerdo a lo anterior $\mathbf{W}$ es simétrica, el criterio E_LE puede escribirse en forma de matriz como se indica a continuación:

$$E_{LE} = tr(\mathbf{XLX}^T) \quad . \tag{7-39}$$

En esta ecuación, $\mathbf{L}$ es la matriz laplecian ponderada del grafo G, definida como

$$\mathbf{L} = \mathbf{W} - \mathbf{D} \quad , \tag{7-40}$$

donde $\mathbf{D}$ es una matriz diagonal con entradas $d_{i,i} = \sum_{j=1}^{N} w_{i,j}$. Para probar la igualdad, basta notar que para un p-incrustación dimensional:

$$E_{LE} = \frac{1}{2}\sum_{i,j=1}^{N} ||\mathbf{x}(i) - \mathbf{x}(j)||_2^2 w_i, j \tag{7-41}$$

$$= \frac{1}{2}\sum_{p=1}^{P}\sum_{i,j=1}^{N}(x_p(i) - x_p(j))^2 w_{i,j} \tag{7-42}$$

$$= \frac{1}{2}\sum_{p=1}^{P}\sum_{i,j=1}^{N}(x_p^2(i) + x_p^2(j) - 2x_p(i)x_p(j))w_{i,j} \tag{7-43}$$

$$= \frac{1}{2}\sum_{p=1}^{P}\left(\sum_{i=1}^{N}(x_p^2(i)d_{i,i} + \sum_{j=1}^{N} x_p^2(j)d_{j,j} - 2\sum_{i,j=1}^{N} x_p(i)x_p(j)w_{i,j}\right) \tag{7-44}$$

$$= \frac{1}{2}\sum_{p=1}^{P} 2\mathbf{f}_p^T(y)\mathbf{D}\mathbf{f}_p(y) - 2\mathbf{f}_p^T(y)\mathbf{W}\mathbf{f}_p(y) \tag{7-45}$$

$$= \frac{1}{2}\sum_{p=1}^{P} \mathbf{f}_p^T(y)\mathbf{L}\mathbf{f}_p(y) = tr(\mathbf{X}\mathbf{L}\mathbf{X}^T) \quad , \tag{7-46}$$

donde $\mathbf{f}_p(y)$ es un N-vector dimensional que da la cota de posición para cada punto incrustado, y $\mathbf{f}_p(y)$ es la transposición del p-th fila de $\mathbf{X}$.

Por cierto, cabe destacar que el cálculo anterior también muestra que $\mathbf{L}$ es positivo semidefinido. Minimizando E_{LE} con respecto a $\mathbf{X}$ bajo la restricción $\mathbf{X}\mathbf{D}\mathbf{X}^T = \mathbf{I}_{P\times P}$ se reduce a resolver el problema del valor propio generalizado $\nearrow\mathbf{D}\mathbf{f} = \mathbf{L}\mathbf{f}$ y en busca de los propios vectores P de $\mathbf{L}$ asociados con los valores propios más pequeños.

A medida que $\mathbf{L}$ es semidefinido simétrico y positivo, todos los valores propios son reales y no inferiores a cero. Esto se puede ver resolviendo el problema de forma incremental, es decir, calculando primero una incrustación unidimensional, luego una bidimensional y así sucesivamente. En este punto, debe notarse que $\nearrow\mathbf{D}\mathbf{f} = \mathbf{L}\mathbf{f}$ posee una solución trivial. De hecho, para $\mathbf{f} = \mathbf{1}_N$ donde $\mathbf{1}_N = [1, ..., 1]^T$, sale a la luz que $\mathbf{W}\mathbf{1}_N = \mathbf{D}\mathbf{1}_N$ y así que $\mathbf{L}\mathbf{1}_N = \mathbf{0}_N$. De ahí $\nearrow_N = 0$ es el valor propio más pequeño de $\mathbf{L}$ y $\mathbf{f}_N(y) = 1$.

Algoritmo (Laplacian Eigenmaps - LE).

---

1. Si los datos consisten en distancias de par, omita el paso 2 y vaya directamente al paso 3.
2. Si los datos consisten de vectores, entonces compute todas las distancias en sentido de los pares.
3. Determinar cualquier K-ary vecindarios o ϵ-ball vecindarios.
4. Construir el gráfico correspondiente y su matriz de adyacencia $\mathbf{A}$.
5. Aplique el kernel de calor (u otro) a los puntos de datos adyacentes y construya la matriz $\mathbf{W}$ como en (7-38).
6. Suma de todas las columnas $\mathbf{W}$ para construir la matriz diagonal $\mathbf{D}$, que consiste en las sumas en filas de $\mathbf{W}$.
7. Cálculo $\mathbf{L}$, el laplaciano de la matriz $\mathbf{W}$:$\mathbf{L} = \mathbf{W} - \mathbf{D}$.
8. Normalizar la matriz laplacica: $\mathbf{L}' = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.
9. Calcule la EVD del Lapón normalizado: $\mathbf{L}' = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.
10. Una incrustación de baja dimensión se obtiene finalmente multiplicando los vectores propios por $\mathbf{D}^{1/2}$, de transponerlas, y mantener a los asociados con el P valores propios más pequeños, excepto el último, que es cero.

7.4. Métodos Basados en Divergencias

7.4.1. Stochastic Neighbor Embedding (SNE)

Incrustación de vecinos estocásticos (Stochastic neighbor embedding - SNE) es uno de los métodos más recientes para la reducción de dimensión [155]. SNE minimiza la divergencia de información D entre dos distribuciones $\boldsymbol{P}_n = [p_{nm}]_{1\leq m\leq N}$ y $\boldsymbol{Q}_n = [q_{nm}]_{1\leq m\leq N}$ asociado con el n-ésimo punto del espacio observado y del espacio latente, respectivamente. Entonces, usando la divergencia dirigida por Kullback-Leibler, D_{KL}, la función objetiva de SNE está en la forma:

$$E_{\mathrm{SNE}}(\boldsymbol{X}) = \sum_{n=1}^{N} \mathrm{D}_{\mathrm{KL}}(\boldsymbol{P}_n \| \boldsymbol{Q}_n) = \sum_{n,m=1}^{N} p_{nm} \log \frac{p_{nm}}{q_{nm}}. \tag{7-47}$$

Definiendo $\delta_{nm} = \|\boldsymbol{y}_n - \boldsymbol{y}_m\|^2$ y $d_{nm} = \|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2$, distribuciones $\boldsymbol{P}_n$ y $\boldsymbol{Q}_n$ pueden ser elegidas como afinidades no simétricas generalizadas y normalizadas en la forma

$$p_{nm} = \frac{\exp\left(-0.5\delta_{nm}^2/\sigma_n^2\right)}{\sum\limits_{n\neq m'} \exp\left(-0.5\delta_{nm'}^2/\sigma_n^2\right)}, \quad \text{and} \quad q_{nm} = \frac{\exp\left(-0.5d_{nm}^2/\pi_n^2\right)}{\sum\limits_{n\neq m'} \exp\left(-0.5d_{nm'}^2/\pi_n^2\right)}, \tag{7-48}$$

con $q_{nn} = 0$ y $p_{nn} = 0$. Se puede lograr una versión simétrica de SNE (SSNE) seleccionando afinidades completamente normalizadas que pueden obtenerse fácilmente modificando ligeramente las ecuaciones. (7-48) y (7-48). En lugar de una suma restringida, todas las entradas deben sumarse en el denominador a fin de hacer cumplir que todas las entradas normalizadas sumen lo siguiente 1, es decir, para garantizar que $\mathbf{1}_N^\top \boldsymbol{Q} \mathbf{1}_N = \mathbf{1}_N^\top \boldsymbol{P} \mathbf{1}_N = 1$.

7.4.2. *t*-SNE

Los métodos basados en SNE sufren de alcanzar espacio latente distorsionado y superpuesto, cuando d es más pequeña que la dimensión intrínseca [166]. Para hacer frente a este problema, se planteó otra variante llamada *t*-SNE y consiste en definir $\boldsymbol{Q}_n$ como una distribución de *t*-Student [151].

Parte IV.

Metodología

8. Metodología y Marco Experimental

8.1. Introducción

El objetivo de la reducción de la dimensión (RD) es, obtener representaciones de datos de baja dimensión en correspondencia a los datos de entrada de alta dimensión, manteniendo la estructura de datos tan bien como sea posible bajo un criterio preestablecido. Alcanzar este objetivo implica que tanto el rendimiento de un sistema de reconocimiento de patrones como la representación inteligible de los datos, pueden mejorarse [167]. Tradicionalmente, los métodos RD se diseñan siguiendo criterios de optimización preestablecidos y parámetros de diseño. Entonces, en su mayoría carecen de las propiedades de interactividad y controlabilidad, siendo características del campo de Visualización de la Información (InfoVis) [168]. InfoVis proporciona interfaces y formas gráficas de representación de datos, que hacen de la información disponible, un elemento más utilizable e inteligible para el usuario. Dicho esto, surge la premisa donde los resultados de la RD pueden ser mejorados aprovechando algunas propiedades de los métodos InfoVis [169] [170]. Siguiendo esta premisa, algunos enfoques han propuesto [9] [171] [12] haciendo uso de la interactividad utilizando interfaces de posición sobre la escala o modelos de interacción geométrica. En general, tales enfoques implementan interesantes modelos interactivos, pero su visualización final carece de la información sobre la estructura de datos del espacio de entrada original, o al menos de una manera fácil de entender y/o visual.

Esta sección se organiza de la siguiente manera: En la sección 8.2, Visualización de datos mediante reducción dimensional. La Sección 8.5 presenta el esquema de visualización interactiva de datos. La configuración experimental del modelo y los resultados basados en similaridad se muestran en las Secciones 8.4.5 y 8.5.2, respectivamente. Finalmente, la Sección 8.7 recoge los elementos usados en el experimento tales como bases de datos y demás insumos de investigación.

8.2. Visualización de Datos Basada en Reducción de la Dimensión

Tal vez, una de las maneras más intuitivas de visualizar datos numéricos es, a través de una representación bidimensional o tridimensional de los datos originales, que se puede representar fácilmente utilizando un gráfico de dispersión. En consecuencia, la reducción de la dimensionalidad surge en correspondencia para que permita alcanzar una representación de datos en baja

dimensión, sobre el cual tanto el rendimiento de la tarea de clasificación se mejora en términos de precisión, como la naturaleza intrínseca de los datos, se representa adecuadamente [172]. Por tanto, cuando se realiza un método de RD, se espera una visualización más realista e inteligible para el usuario [167].

Técnicamente, el objetivo de la reducción de la dimensión es integrar una matriz de datos de alta dimensión. $Y = [y_i]_{1 \leq i \leq N}$ tal que $y_i \in \Re^{N \times N}$ en una matriz de datos latentes de baja dimensión $X = [x_i]_{1 \leq i \leq N}$ posterior $x_i \in \Re^{N \times N}$, donde $d < D$. Fig. (B.2) representa un caso del cual un colector, llamado ***estructura esférica artificial en 3D***, se encuentra representado en un espacio de menor dimensión **(2D)**, que se asemeja a una versión desplegada del colector original.

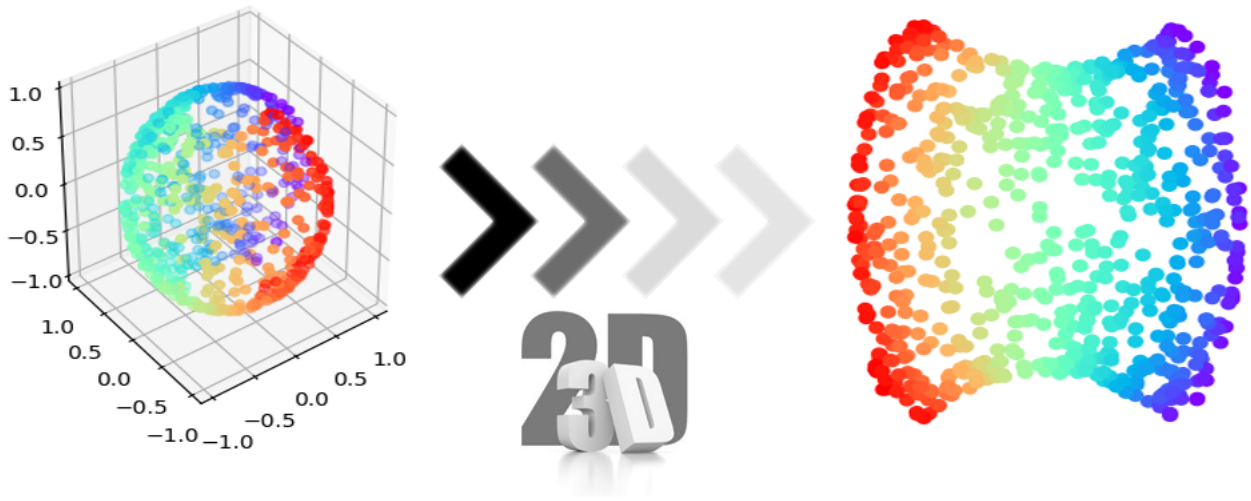


Figura 8-1.: Estructura Esférica Artificial, representación dimensional de (3D) a (2D).

Los enfoques clásicos de la RD, apuntan a preservar la varianza (análisis de componentes principales - PCA) o la distancia (escalamiento multidimensional clásico - CMDS) [168]. Hoy día, los métodos más desarrollados y recientes, apuntan a preservar la topología de los datos. Tal topología puede ser representada por un grafo basado en datos, construido como un grafo no dirigido y ponderado, en el cual los puntos de datos representan los nodos, y una matriz de similitud no negativa (también afinidad) sostiene los pesos de las aristas a pares. Esta representación se explota tanto con métodos espectrales como basados en divergencias. Por un lado, para los enfoques espectrales, la matriz de similitud puede representar el factor de ponderación para distancias a pares, como sucede en Laplacian EigenMaps (LE) [173]. Asimismo, utilizando una matriz de similitudes asimétricas y centrándose en la estructura local de los datos, del cual surgió el método de Locally Linear Embeding (LLE) [174]. Por otro lado, una vez normalizada, la matriz de similitud también puede representar distribuciones de probabilidad, al igual que los métodos basados en divergencias

como Stochastic Neighbor Embeding (SNE) y t-distributed SNE (t-SNE) [167].

8.3. Modelo (DataVisSim) Data-Visualization-Similarity

El modelo (DataVisSim "Data-Visualization-Similarity"), introduce un nuevo enfoque de visualización utilizando una mezcla interactiva de representaciones de datos resultantes de los métodos RD. Después de aplicar los métodos RD en los datos de entrada, se obtiene un conjunto de espacios de representación de menor dimensión. Particularmente, la mezcla se hace a través de una suma ponderada. Para proporcionar a los usuarios un sentido de la estructura de datos, además se implementó una visualización basada en datos, además del gráfico de dispersión convencional. Esta visualización captura la estructura de los datos de entrada, utilizando una matriz de similitud (también, matriz de afinidad de la teoría de grafos), que capta el grado de similitud o afinidad entre cada par de puntos de datos. La visualización consiste en trazar aristas (edges) entre los puntos de datos que presentan el mayor valor de similitud. Además, para proporcionar un mayor sentido de interactividad, el usuario puede controlar el número de aristas mediante un parámetro-variable, que funciona como una barra deslizante dentro de una interfaz. Por diseño, la afinidad se selecciona como gaussiana para que se tenga en cuenta la estructura de los puntos vecinos locales. Particularmente, los espacios de baja dimensión se obtienen por métodos de última generación como: Classical Multidimensional Scaling (CMDS) [168], Laplacian Eigenmaps (LE), Locally Linear Embedding (LLE) [175], Stochastic Neighbor Embedding (SNE), and t-Student-distributed- SNE (t-SNE) [167] [173]. Para realizar la mezcla, el usuario puede ajustar los factores de ponderación recogiendo valores de una interfaz similar a una barra ecualizadora. Para probar el enfoque de visualización, se utiliza un conjunto de datos de estructura esférica artificial en 3D. La calidad de los espacios de representación resultantes se cuantifica mediante una versión a escala de la tasa media de concordancia entre K-ary neighborhoods [174]. La mezcla propuesta puede representar cada uno de los enfoques de reducción de la dimensión y ayuda a los usuarios a encontrar una representación adecuada de los datos de entrada dentro de una interfaz visual y amigable para el usuario.

8.3.1. Selección del Algoritmo mediante Visualización e Interactividad en RD

Un componente importante en esta investigación fue *"Visualización Usando Reducción de Dimensión e Interactividad"*, que se introdujo en el Estado del Arte en la (sección 6.5). Dicho componente posee tres enfoques, del cual se hizo especial hincapié en *"La idoneidad en selección de algoritmos"* (ver sección 6.5.3 del Estado del Arte) [147]. Tal perspectiva identifica la necesidad del aumento de las capacidades humanas mediante la visión, debido al amplio ancho de banda del campo visual humano [40], para ser potenciado mediante procesos como la interpolación general

con sus vertientes aplicadas como (Interpolación de Movimiento e Interpolación de Forma) [147], con la finalidad de representar los factores de interactividad en las morfologías visuales, tal como *INFOVIZ*, que para efectos de visualización de la información y *RD-Interactive* [12], focalizan esfuerzos en tener un amplio espectro de observación, en los movimientos intermedios que permitan el estudio profundo de los procesos intrínsecos de tales comportamientos, en el ejercicio del campo de la reducción de la dimensión interactiva [9].

8.3.2. Interpolación General e Interactividad

El proceso de interpolación en el sub-campo matemático, obtiene nuevos datos teniendo en cuenta el conocimiento previo de un conjunto de datos. Básicamente permite construir una función más compleja o detallada a partir de un muestreo simple, el cual resulta en el ajuste de esta nueva función compleja, con la finalidad de poder determinar detalladamente el desarrollo de la ejecución del proceso a interpolar [176].

En computación gráfica el concepto se hace más simple, al entender que hay dos estados denominados *"Inicio y Fin"*, dicha interpolación actúa bajo el principio de construir los N movimientos intermedios entre estos dos estados, mediante los ejercicios de inferencia que permite establecer los puntos. Tal concepto, se hace presente en la interpolación lineal famosa-mente aplicada a tareas de computación gráfica [177]. Al desarrollar los N movimientos intermedios, gráficamente se puede determinar un gran conjunto de observaciones que describe el proceso con un alto detalle para su procesamiento posterior, que en mayoría de casos puede ser material relevante en tareas de interactividad [178]. El concepto de interactividad, posee requisitos básicos de funcionalidad como la disposición de un conjunto grande de observaciones el cual, la interpolación realiza un gran aporte, debido a la capacidad de brindar dicho conjunto de observaciones, gracias a los diferentes métodos de interpolación tales como: Interpolación bilineal, Interpolación lineal, Interpolación multivariable, Interpolación polinómica, Interpolación polinómica de Hermite, Interpolación polinómica de Lagrange, Interpolación polinómica de Newton, Interpolación por el vecino más cercano y finalmente la Interpolación trigonométrica como parte del grupo de las mas usadas, que finalmente la selección de alguna de ellas soluciona y apoya en gran medida al proceso de interactividad, mediante el cual utiliza los valores de la interpolación, mediante eventos de selección de parámetros, interactúan con el usuario [179].

Hay varios tipos de interpolación como se menciona anteriormente, del cual para efectos de esta investigación se utilizó la interpolación lineal. En virtud de ello tal método, permite estudiar los movimientos intermedios en la matriz de similitud del modelo aquí expuesto (DataVisSim), mediante el diseño de ecualizadores paramétricos interactivos, para establecer de una manera intuitiva la conexidad representada por los puntos pares de cada nodo, así como también los eventos polimorfistas que surgen cuando se pueden referenciar cada uno de los valores de la reducción de la dimensión, para evidenciar todos los cambios de la morfología visual en las salidas gráficas, el

cual se puede observar de una manera estructurada e intuitiva para su posterior comparación visual o estudio cualitativo correspondiente.

8.3.3. Interpolación de Movimiento

La interpolación de movimiento es una aplicación de la *Interpolación Lineal*, que subyace dentro del paradigma de la computación gráfica, famosa en tareas de computación gráfica y procesos de animación digital [180].

Figura 8-2.: La interpolación de movimiento se aplica a la matriz de afinidad, que para efectos de entendimiento intuitivo, se desarrolló un slider que controla el grado de conexidad mediante la distancia de pares (pairwise) entre cada nodo.

Los aportes para ésta investigación surgen dentro del proceso de exploración visual desarrollado en el modelo (DataVisSim), el cual permite evidenciar de una manera intuitiva uno de los componentes de la interfaz aquí desarrollada, que pretende mostrar de una manera intuitiva el porcentaje de afinidad de cada vértice (nodo), del grafo de representación (Estructura Esférica Artificial en 3D), para efectos de estudiar la estructura del espacio en alta dimensión.

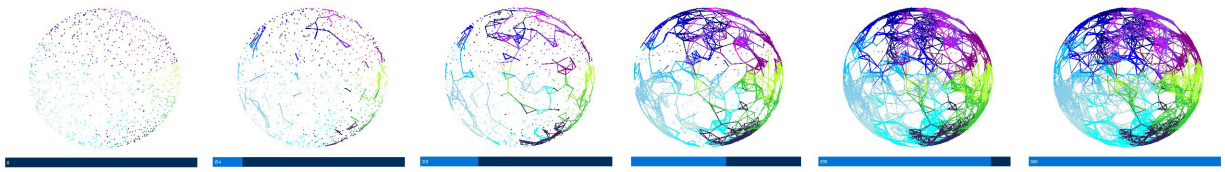


Figura 8-3.: Aplicación de interpolación de movimiento, dentro del proceso de relación Nodo-Aristas, mediante distancia de pares (PairWise), para la representación de la matriz de afinidad. Dicha representación hace referencia a la topología de los datos en un espacio de alta dimensión.

8.3.4. Interpolación de Forma

La interpolación de forma es una aplicación de la *Interpolación Multivariable*. En consecuencia de ello, centra especial esfuerzos en la representación visual de transformaciones de elementos u objetos de tipo primitivas, según la computación gráfica (Puntos, Líneas, Ángulos, Cuadrados, Círculos, Triángulos) [181].

Figura 8-4.: Tal concepto de interpolación de forma, se aplica a la representación de datos de baja dimensión. Para efectos de entendimiento intuitivo, se desarrolló el modelo de ecualizador para manipular la (Reducción de la Dimensión Interactivamente). El objetivo fundamental es realizar una mezcla ponderada de los métodos de RD seleccionados, que obtiene finalmente esta representación en un espacio de baja dimensión.

Los aportes para esta investigación, se conciben dentro de las necesidades de maximizar la exploración visual, de aquellos resultados inmersos en las morfologías visuales en espacios de baja dimensión. Este aporte permite ver de una manera intuitiva, toda aquella información intrínseca sobre el desarrollo de las transformaciones de la estructura, desde su momento inicial, o sea en un espacio de alta dimensión, hasta su transformación final en la representación gráfica que corresponde a una morfología visual en un espacio de baja dimensión.

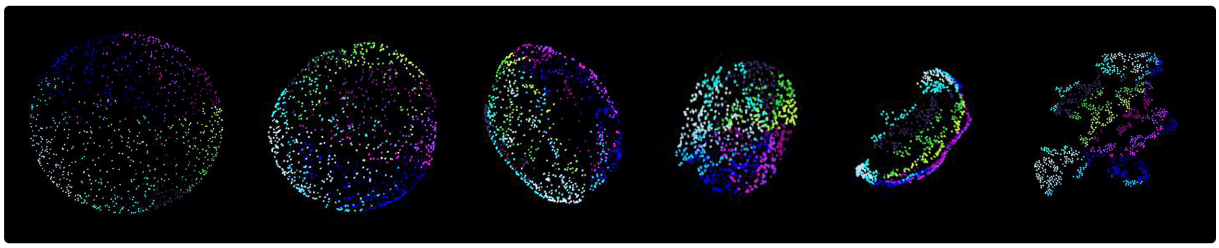


Figura 8-5.: La interpolación de forma, se aplica a la representación de datos de baja dimensión. Para efectos de entendimiento intuitivo, se desarrolló el modelo de ecualizador para manipular la (Reducción de la Dimensión Interactivamente).

8.3.5. Análisis Exploratorio Científico con DataVisSim

Un aporte muy importante en esta investigación, es la posibilidad de interactuar con una morfología visual mediante métodos de reducción de la dimensión para labores de *"Análisis Exploratorio para el Descubrimiento Científico"*, que se introdujo en el contexto de esta investigación (ver sección 5.2.6). Tal proceso permite refinar, depurar o ampliar los algoritmos para entender intuitivamente como se ven afectados por los eventos de cambios de parámetros. Claramente el marco conceptual de esta investigación está dirigido a todos aquellos que necesitan de un sistema superior, el cual ayude a determinar cuál de los múltiples métodos de reducción de la dimensión seleccionados es el más idóneo sin ser experto en el tema, así como también, determinar la configuración ideal de la mezcla entre ellos, para hallar su salida el cual, pueda satisfacer el análisis exploratorio para el descubrimiento científico, cuya labor principal es maximizar las capacidades humanas del usuario, para generar y comprobar hipótesis.

8.3.6. Criterios de Selección en Algoritmos: CMDS, LLE, LE, SNE, TSNE

Los métodos de reducción de la dimensión, se pueden agrupar de la siguiente manera: Métodos Espectrales basados en Similitudes, Disimilitudes, Kernel, Distancias. Métodos Estocásticos basados en Divergencias. Métodos Heurísticos basados en arquitecturas de Redes Neuronales Artificiales, Bayesianas, Recurrentes y Profundas.

Según autores con gran renombre dentro del estado del arte, como: John A. Lee, Michel Verleysen en su libro *Nonlinear Dimensionality Reduction* [137], así como también Joshua B. Tenenbaum, Vin de Silva, John C. Langford en su libro *A Global Geometric Framework for Nonlinear Dimensionality Reduction* [182], Geoffrey Hinton, Sam Roweis en su libro *Unsupervised learning foundations of neural computation* [183], afirman que los métodos espectrales basados en similitudes como; LLE [88], LE [89]. Espectrales basados en disimilitudes como CMDS [102] y Métodos estocásticos basados en divergencias como SNE [116], son la base fundamental de casi, todos los métodos existentes del paradigma no lineal.

Básicamente la gran mayoría de métodos del estado del arte actual, son mejoras (tal como se expone en la sección 6), de la presente investigación en RD, dicha afirmación realiza especial énfasis en las similitudes compartidas dentro del proceso de optimización del cálculo matricial y la representación de sus colectores en nuevas formas y modelos que definen notables progresos. De este modo, se enfocan principalmente en el desarrollo de representaciones de datos, para ser proyectados a nuevas y optimizadas representaciones de baja dimensión. Los métodos seleccionados en esta investigación representa la generalidad de la mayoría de metodologías que están clasificadas bajo la taxonomía de la reducción de la dimensión (ver Estado del Arte, sección 6.4.3), expuesta

en el estado del arte actual, que corresponde a fecha del año 2017 de esta tesis de maestría.

8.4. Morfología Visual de Representación

Las morfologías visuales o Modismos se construyen bajo las metodologías de Canales y Marcas según el framework de Tamara Munzner en su libro de *Visualization Analysis and Design: Principals, Techniques, and Practice* [1], el cual poseen métricas que pueden medir factores de expresividad, así como factores de efectividad. En esta investigación se realizó un estudio de las marcas y canales más expresivos y efectivos, en términos de la interacción entre humano ordenador, para satisfacer la tarea de poder representar la reducción de la dimensión interactiva de manera intuitiva. A continuación se exponen los criterios de selección, dentro de los aspectos de los Canales y Marcas utilizados en los modismos que hacen parte del ecosistema de la metodología de visualización creada mediante el desarrollo de DataVisSim.

8.4.1. Clases de Expresividad y Clasificación de Efectividad

Según Tamara Munzner, en su trabajo de investigación ”visualization analysis and design”, afirma que; las representaciones visuales con bases de codificación como primitivas geométricas, se les denomina Marcas, el cual permiten establecer un modelo que se desarrollara de tal manera que otros canales, tal como el visual, se encargará de controlar su apariencia. A continuación se mencionan algunos de los más usados:
En la ciencia de visualización, se recomienda comprender acerca de las marcas y los canales (ver en la figura (**8-6**)). Este enfoque, propone las bases del análisis y construcción de codificaciones visuales. Hablando en término de codificaciones visuales, el espacio de diseño se describe como una combinación ortogonal de dos conceptos tales como: elementos gráficos llamados marcas y canales visuales para controlar su apariencia, de este modo se observa que las codificaciones visuales complejas también aplican para ser analizados en términos de sus marcas y estructura de canales.

8.4.2. Expresividad y Efectividad en Marcas Utilizadas

En materia de codificación visual, hay un elemento que resalta de todos y posee importante protagonismo por ser la base fundamental de las morfologías visuales, a este elemento se denomina las **Marcas**. Dicho concepto está compuesto por todos los componentes tales como: las famosas *Primitivas Geométricas*. Este concepto posee aquellas geometrías como (Puntos, Líneas, Áreas), que permiten forjar la estructura de la morfología visual del modismo.

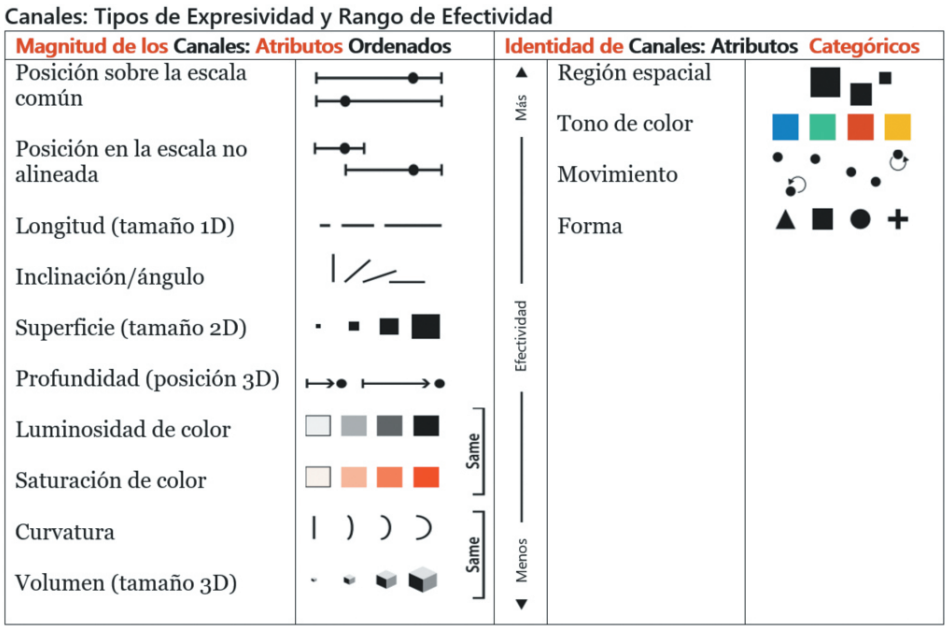


Figura 8-6.: La eficacia de los canales que modifican el aspecto de las marcas depende de la expresividad de canales con atributos codificados. (Tamara Munzner, libro "visualization analysis and design" 2014 [1])

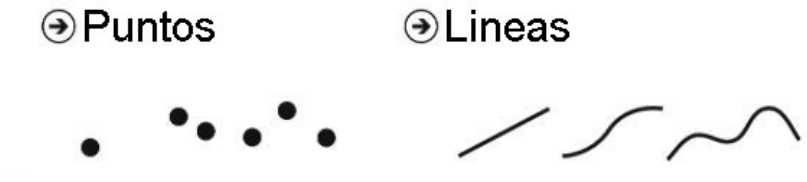


Figura 8-7.: Primitivas Geométricas denominadas Marcas. (Tamara Munzner, libro "visualization analysis and design" 2014 [1])

En esta investigación se utilizó dos de las Marcas con mayor grado de expresividad y efectividad como: **Líneas**, que para efectos de diseño, están representados en los modelos de barras de ecualizadores, utilizados por pertenecer al grupo de las primitivas geométricas con mayor indice de efectividad y expresividad en el ser humano, según *(**Tamara Munzner** - Universidad de Stanford en su investigación Visualization Analysis and Design)* [1] y **Puntos** como el gráfico de dispersión base, que representa cada uno de los datos, en la representación de la base de datos (Estructura Esférica Artificial en 3D).

8.4.3. Expresividad y Efectividad en Canales Utilizados

En la teoría de codificación visual, los **canales** corresponden al medio de transformación, que nos permite modificar la apariencia de una Marca (Primitivas Geométricas), el cual posee tipos de expresividad, rangos de efectividad, tal como se muestra en la figura **8-6**. Tales características, se agrupan en diferentes categorías como: Magnitud de los Canales, Identidad de los Canales, con sus correspondientes Atributos Ordenados y Atributos Categóricos. Para el ejercicio de esta investigación se tuvo en cuenta las *Clases de Expresividad y Clasificación de Efectividad*, el cual presenta los canales más expresivos y efectivos para desarrollar codificaciones visuales. Para el diseño de la interface DataVisSim, se utilizó el canal *Posición sobre la escala* catalogado como el más efectivo según el rango de efectividad y expresividad (ver figura **8-6**), para el diseño del modelo de ecualizadores, que permite al usuario evaluar una serie de valores correspondientes al espectro de la reducción de la dimensión, mediante una escala que permite adaptar de 0% a 100% en porcentaje de aplicación del método de RD.

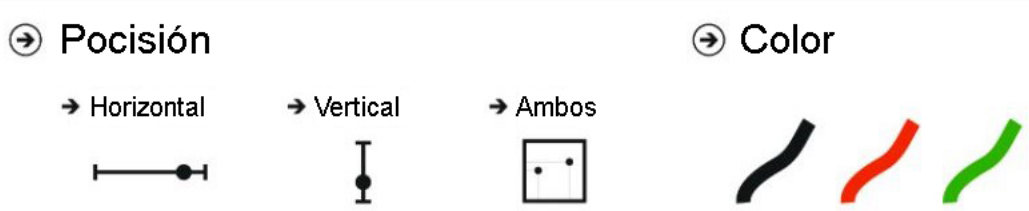


Figura 8-8.: Primitivas Geométricas denominadas Marcas. (Tamara Munzner, libro ”visualization analysis and design” 2014 [1])

En el diseño de la interfaz se utilizó el canal del color. Debido a su alto rendimiento de efectividad y expresividad según el rango de efectividad y expresividad (ver figura **8-6**), el cual representa las (clases o grupos) que se implementaron en el modelo de base de datos (Estructura Esférica Artificial en 3D - ver figura **8-14**), el cual permite un alto sentido de orientación de identidad categórico, tal cual como se muestra en la figura **8-6**.

Por último el canal del movimiento fue utilizado (ver figura **8-6**), para realizar las recreaciones del componente de la matriz de afinidad que se desea estudiar, respecto a la estructura de los datos en espacios de alta dimensión. El cual permite ver claramente el conjunto de movimientos a nivel de conexidad de los N vértices con las N aristas.

8.4.4. Analogía del Modelo de Reducción de la Dimensión Interactiva

En esta sección se aborda una explicación intuitiva referente al modelo DataVisSimm. El desarrollo de este proceso consta de una serie de pasos que llegan a final término con una representación de dimensión menor, respecto al espacio original de datos. Para dicho objetivo se expone; una analogía, con el único propósito de crear una imagen clara e intuitiva al investigador o usuario que desee replicar el experimento, mediante el modelo de ecualizadores propuesto.

Suponga que tiene una tarea definida en temas relacionados con: Minería de Datos, Aprendizaje de Máquina, Analítica de Datos o Exploración de Datos, del cual; un equipo de trabajo pretende seleccionar alguno de los métodos de reducción de la dimensión, tales como: *(CMDS, LLE, LE, SNE, t-SNE)* o en el mejor de los casos impartir un análisis exploratorio científico, que pueda realizar una mezcla entre ellos y explotar las mejores propiedades de cada uno de estos métodos y así crear sus propias configuraciones, que mejor representen la información original, para ser asociada a un espacio de baja dimensión.

Este proceso se realiza mediante una mezcla interactiva en tiempo real, mediante el cual; se puede dar nociones fuertes del proceso, para un mejor entendimiento la topología de estos datos. Dicho proceso permite evidenciar desde su inicio, como una estructura en un espacio de alta dimensión se transforma, para conducir al investigador por todas las fases de transformación de estos datos, permitiendo el paso de parámetros el cual brinda la posibilidad de verlo y personalizarlo todo, hasta llegar intuitivamente a encontrar la configuración deseada, el cual satisface mejor la tarea a realizar.

El proceso más interesante es la ***Mezcla Interactiva de Reducción de la Dimensión***, puesto que permite crear nuevos métodos, partiendo de la explotación de las exorbitantes combinaciones que se pueden lograr con los métodos seleccionados. Identificando esta necesidad se imparte la analogía de la siguiente manera:

Analogía: Suponga que se tienen los ingredientes ideales para crear ***(la fórmula perfecta de Reducción de la dimensión)*** tal como se muestra en la figura **8-9**, entonces se debe definir cuál es la medida correcta o ideal en los ingredientes, para que dicho resultado ***sea el mejor en la resolución del problema***, se buscará en los parámetros aquellos valores de forma intuitiva, la mejor configuración utilizando el canal visual.

A priori, se asume que el usuario no tiene experiencia en este procedimiento, por tal razón no hay conocimiento que garantice lo asertivo de sus medidas respecto a cada ingrediente. Para cumplir con su cometido es necesario empezar a realizar las pruebas sobre aquellos ingredientes. Este proceso permite determinar de una manera fácil, cuál será el mejor porcentaje de cada ingrediente para realizar la mezcla que resulta, en el mejor método de Reducción de la Dimensión para el problema a tratar.

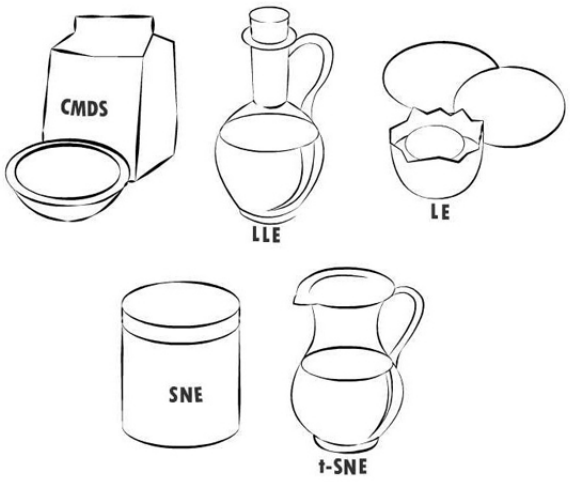


Figura 8-9.: Los ingredientes para la Mezcla de la Reducción de la Dimensión Interactiva.

Mediante un proceso de prueba, el cual ajusta los porcentajes de cada ingrediente, se tiene la medida que se acerca a dicha consistencia y sabor como se muestra en la figura **8-10**. De esta manera surge la mezcla perfecta que resultará en una configuración ideal para la construcción de la Reducción de la Dimensión.

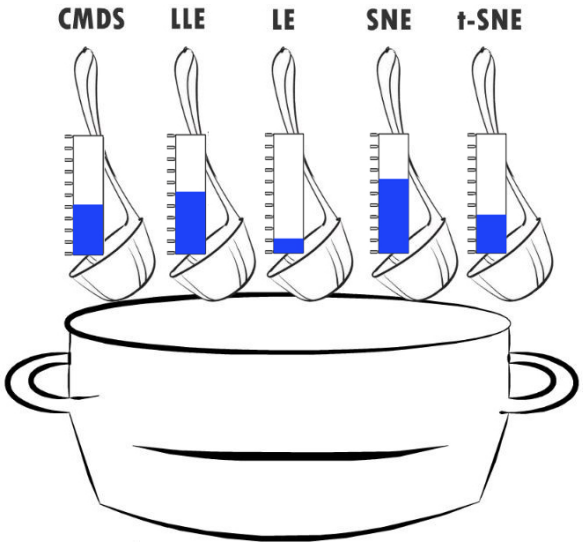


Figura 8-10.: El porcentaje de configuración de las porciones o parámetros en los ingredientes para la Mezcla de la Reducción de la Dimensión Interactiva.

Este proceso enriquece en gran medida los resultados, ya que mediante las combinaciones que vayan resultando se puede ***Ver y Estudiar*** cada uno de los movimientos internos, que de otra man-

era difícilmente se podría evidenciar. Este proceso da la posibilidad de probar hipótesis, refinar, y ajustar las N configuraciones posibles para encontrar una solución de buena calidad.

Esta analogía pretende demostrar, que la inmersión en el proceso, favorece positivamente en la posibilidad de interactuar de manera intuitiva en los parámetros de configuración del modelo ecualizador, sobre la estructura de datos original. Tal proceso, garantiza los mejores resultados de una manera fácil, sin necesidad de un experto, el cual ayude a la interpretación. Esto se debe a las salidas, que están diseñadas bajo los canales y marcas más efectivos, del cual apoya notablemente, al aumento de las capacidades en la correcta toma de las decisiones.

8.4.5. Modelo Interactivo

Uno de los objetivos fundamentales de esta investigación, fue lograr el factor de interactivo. Esto permite al usuario, tener las herramientas de exploración necesarias para apoyar al aumento de las capacidades analíticas, en la correcta selección de la configuración ideal, para apoyar el proceso de mezcla interactiva de reducción de la dimensión.

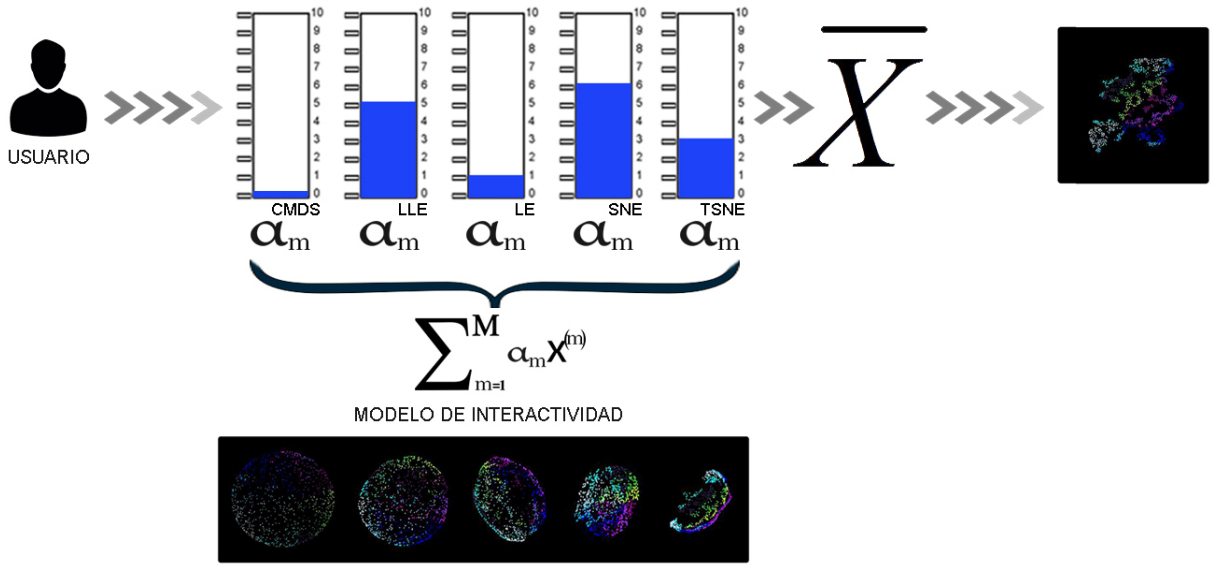


Figura 8-11.: El porcentaje de configuración de los parámetros en los ingredientes, para la Mezcla de la Reducción de la Dimensión Interactiva.

Para alcanzar el objetivo de la interactividad, los valores de cada α_m necesarios para calcular $\bar{X}$ según ecuación (8-1), deben ser definidos por los usuarios utilizando una interfaz similar a una barra ecualizadora, tal como se muestra en la figura **8-11**. Se creó un entorno intuitivo y de usuario amigable para los factores de ponderación, el cual se pueden introducir fácilmente con sólo recoger los valores de las barras. Para proporcionar vistas rápidas del espacio de representación resultante, tan pronto como se recoge un punto, los puntos restantes se completan automáticamente siguiendo

una función de probabilidad de densidad uniforme. Lo mismo se hace si se realiza la selección de más de un valor.

8.5. Esquema Interactivo de Visualización de Datos

El enfoque de visualización propuesto, aquí llamado DataVisSim, involucra tres etapas principales: mezcla de resultados de RD, interacción y visualización, como se muestra en el diagrama de bloques de la figura (**8-13**). Una de las contribuciones más importantes de este trabajo es, la información sobre la estructura del espacio de alta dimensión de entrada, que se suma a la representación visual final, utilizando un esquema basado en la similitud por pareja.

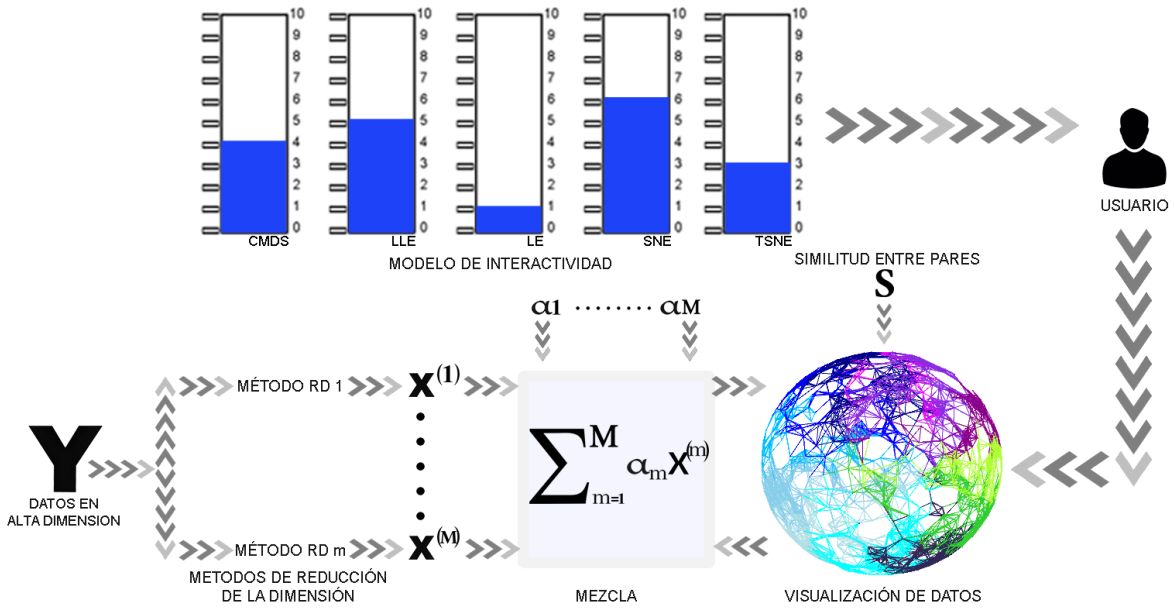


Figura 8-12.: Diagrama de bloques de la visualización interactiva de datos propuesta utilizando la reducción de la dimensión y representaciones basadas en similitudes (DataVisSim).

En términos generales, funciona de la siguiente manera: primero realiza una mezcla de espacios de representación de baja dimensión resultantes aprovechando las implementaciones convencionales de los métodos tradicionales de RD. La interacción se proporciona a través de una interfaz que permite al usuario introducir dinámicamente los factores de ponderación para la mezcla antes mencionada. Para la visualización, se utiliza un enfoque novedoso basado en la similitud.

8.5.1. La Mezcla, desde la Cosmovisión Matemática

Supongamos que la matriz de entrada Y se reduce utilizando M diferentes métodos RD, obteniendo un conjunto de representaciones de menor dimensión: $\left\{X^{(1)}, ..., X^{(M)}\right\}$. Aquí se propone realizar una suma ponderada en la forma:

$$X = \sum_{m=1}^{M} \alpha_m X^{(m)} \tag{8-1}$$

donde $\{a_1, ..., a_M\}$ son los factores de ponderación. Para que la selección de los factores de ponderación sea intuitiva, se utilizan valores probabilísticos de manera que $0 \leq \alpha_m \leq 1 \quad y \quad \sum_{m=1}^{M} \alpha_m = 1$, por lo tanto todas las matrices $X^{(m)}$ deben ser normalizadas para confiar en una hiperesfera de proporciones.

8.5.2. Visualización Basada en Similitudes

El método más utilizado para visualizar datos bidimensionales o tridimensionales es la gráfica de dispersión, teniendo en cuenta que la Marca (Primitivas Geométricas - ”Puntos”) hacen parte del grupo con mayor efectividad y expresividad según el FrameWork de (**Tamara Munzner** - Universidad de Stanford en su investigación Visualization Analysis and Design) [1]. En esta tesis de maestría, se introduce un enfoque de visualización basado en la similitud con el objetivo de proporcionar una pista visual sobre la estructura de la matriz de datos de entrada de alta dimensión Y en el diagrama de dispersión de su representación en un espacio de baja dimensión $\bar{X}$.

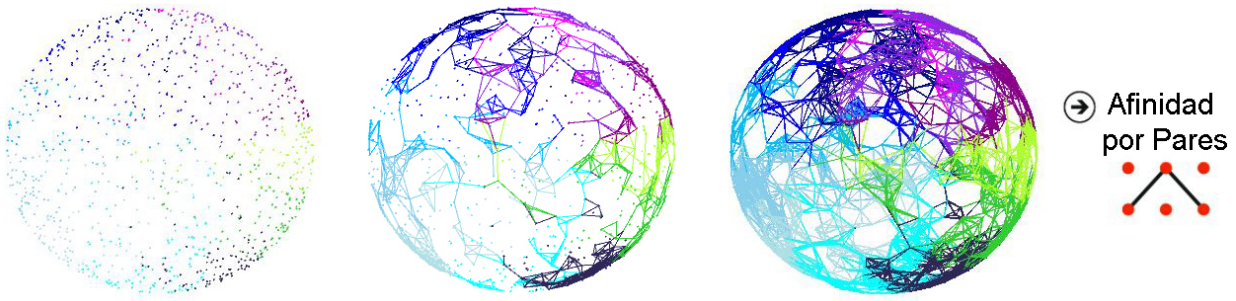


Figura 8-13.: Representación de datos en un espacio de alta dimensión mediante la estructura esférica artificial, el cual esboza la noción de la estructura de datos a tratar, mediante los métodos de la reducción de la dimensión. Vale la pena resaltar el modelo de conexidad entre los puntos (Nodos) y las líneas (Aristas), como Afinidad por Pares. El cual permite ver de una manera intuitiva la transformación final, respecto la estructura original de los datos de entrada.

Para ello, se utiliza una matriz de similitudes por pares $s \in \Re^{N \times N}$, tal que $S = \left[s_{ij}\right]$. En términos de teoría de grafos, las entradas s_{ij} define la similitud o afinidad entre el i-ésimo y j-ésimo punto

de datos de Y. De esta manera, se mantiene la estructura del espacio de entrada original de una manera topológica, específicamente en términos de relaciones de pareja.

Para fines de visualización, esta similitud se utiliza para definir gráficamente la relación entre los puntos de datos mediante el trazado de aristas. Para controlar la cantidad de aristas y realizar representaciones visuales atractivas, el valor de s_{ij} se ve restringida porque $s_{ij} > s_{max}$ más s_{max} usuarios. En otras palabras, el enfoque de visualización consiste en construir un grafo con valores de afinidad limitados.

8.6. Marco Experimental

Base de Datos: Para evaluar visualmente el rendimiento del enfoque DataVisSim, se utiliza una estructura esférica artificial. ($N = 1500 datapoints and D = 3$), como se muestra en Fig. (**8-14**).

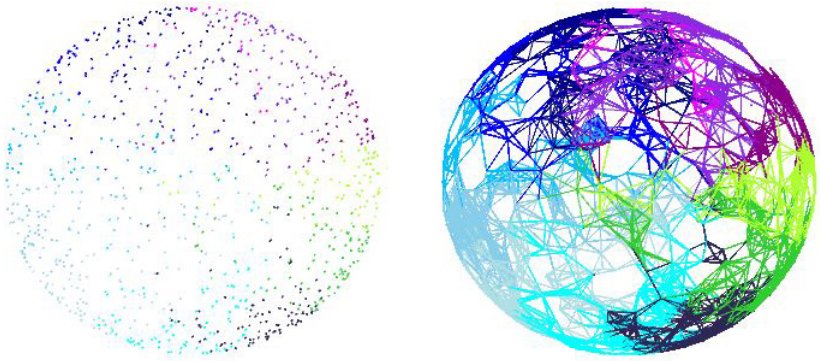


Figura 8-14.: Estructura Esférica Artificial, para representación de topología de datos en espacios de alta dimensión

Ajustes de Parámetros y Métodos: Con el fin de capturar la estructura local para la visualización, es decir, los puntos de datos que son vecinos, se utilizó la similitud gaussiana dada por: $s_{ij} = exp\left(-0.5\left\|y_i - y_j\right\|^2 / \sigma^2\right)$. El parámetro σ es un valor de ancho de banda establecido como 0.1, siendo el 10% de la relación de hiperesfera (aplicable una vez que las matrices son normalizadas como se discute en la Sección (8.5.1). Para realizar la reducción de la dimensión se considera $M = 5$ métodos RD, es decir: CMDS, LE, LLE, SNE y t-SNE. Todos ellos están destinados a obtener espacios en dimensiones $d = 2$.

8.7. Evaluación de la Calidad de la Reducción de la Dimensionalidad

Criterios por orden de clasificación: Para cuantificar el rendimiento de los métodos estudiados, el método $R_{NX}(K)$ creado por [184] se utiliza, dentro del intervalo $[0, 1]$. Desde $R_{NX}(K)$ se calcula en cada valor de perplejidad desde 2 hasta $[N - 1]$, se puede obtener un indicador numérico del rendimiento global calculando su área bajo la curva (AUC). La AUC evalúa la calidad de la reducción dimensional a todas las escalas, con los pesos más adecuados.

Parte V.

Experimentos y Resultados

9. Resultados Experimentales

9.1. Resultados

La importancia de la aplicación de esta métrica, consiste en brindar el beneficio de evidenciar en forma visual, la comparación de cada método con las mezclas resultantes del modelo DataVisSim, el cual evalúa la calidad de todos aquellos aspectos de los métodos que resultan en la preservación de la topología de los datos. Tal comparación de representaciones resultantes de espacios en menor dimensión con $R_{NX}(K)$, para cada valor de perplejidad de 2 a $N-1$, se obtiene el rendimiento global mediante un valor cuantificable que obtiene el área bajo la curva (AUC).

La noción que se entrega bajo el desarrollo del proceso de este método, es la posibilidad de medir el grado de preservación de las estructuras o colectores mapeados, que en este caso es ***(estructura esférica artificial en 3D)***. Se afirma que, mientras los trazados demuestren formas asimétricas pronunciadas al costado derecho, indica que la calidad de esta representación satisface la preservación de la topología de datos, en forma global. De igual manera, sucede cuando el solapamiento del trazado, tienen inclinación opuesta, esto indica que la calidad tiende a salvaguardar la topología local de los datos.

EL enfoque aquí propuesto, realiza especial énfasis en la preservación de la antropología global de los datos, debido a la naturaleza global del colector seleccionado para el experimento ***Estructura Esférica Artificial en 3D***, otro aspecto interesante es la capacidad de los métodos seleccionados, que poseen cualidades de orden espectral y divergente que tienden a obtener mejores representaciones en estructuras globales. Tal afirmación será demostrada a continuación mediante el análisis de las curvas de calidad correspondientes a cada mezcla, dado que demuestra mejores rendimientos cuando posee estas características en mención.

Para el mejoramiento de dicho enfoque, se realiza una mezcla interactiva de métodos de reducción de la dimensión, mediante el cual se evidencia de forma intuitiva, la mejor configuración que permite al modelo de ecualizadores el control intuitivo mediante el factor humano. La ventaja de las mezclas interactivas de reducción de la dimensión es, poder integrarse a este tipo de métricas y dar un diagnóstico del colector para definir si su estructura es de orden local o global. Otro factor imprescindible en esta cualidad, radica en establecer el porcentaje de aplicación cada método para obtener mejores resultados en la mezcla dimensional de representación de menor, teniendo como prior la definición de la morfología de datos del colector a estudiar, ya sea de orden global o local

y el performance o rendimiento de cada método basado en las estructuras globales y locales para explotar al máximo las propiedades de cada método.

En la figura (**9-1**) se puede ver cinco tipos de representaciones en baja dimensión, de los Métodos espectrales basados en similitudes como; LLE [88], LE [89]. Espectrales basados en disimilitudes como CMDS [102] y Métodos estocásticos basados en divergencias como SNE [116]. Los resultados experimentales se obtienen utilizando $R_{NX}(K)$ como indicador de calidad. El $R_{NX}(K)$ permite la comparación de cuatro mezclas diferentes de métodos de DR y cinco métodos de DR para determinar la mejor combinación.

Un hecho interesante encontrado en las curvas de calidad, es la mezcla de los métodos de DR, el cual muestra un área mayor bajo la curva que algunos métodos considerados, este implica que los factores de ganancia en la curva de calidad, deja traslucir que algunos métodos se comportan mejor con este tipo de representaciones de datos artificiales y otros sencillamente funcionan mejor con datos reales.

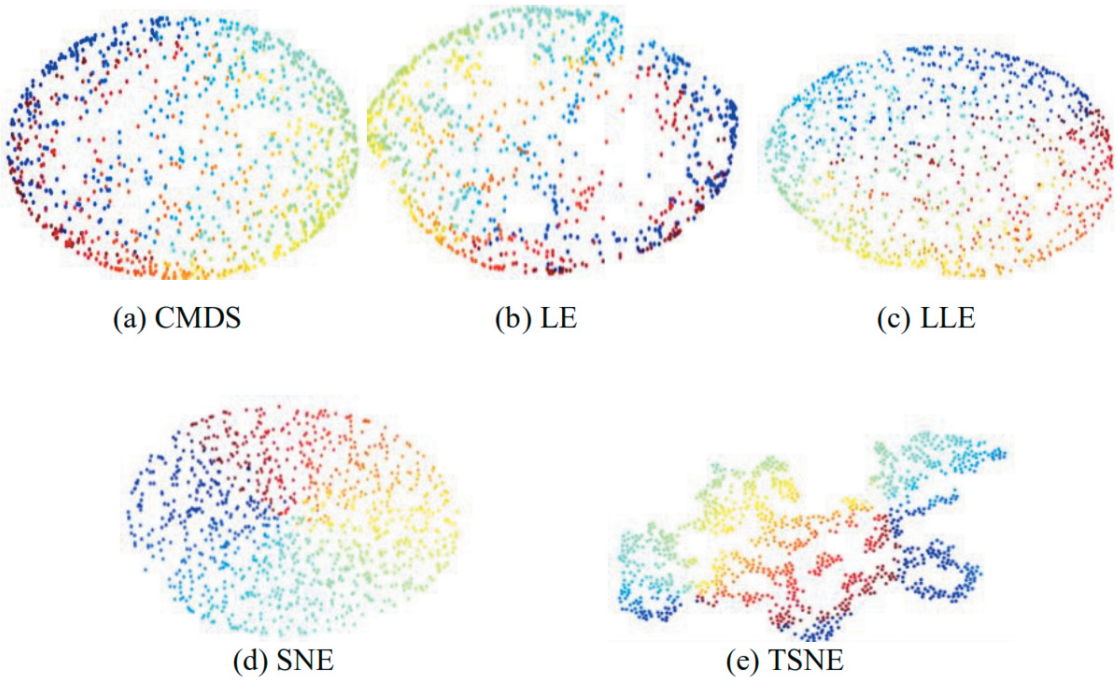


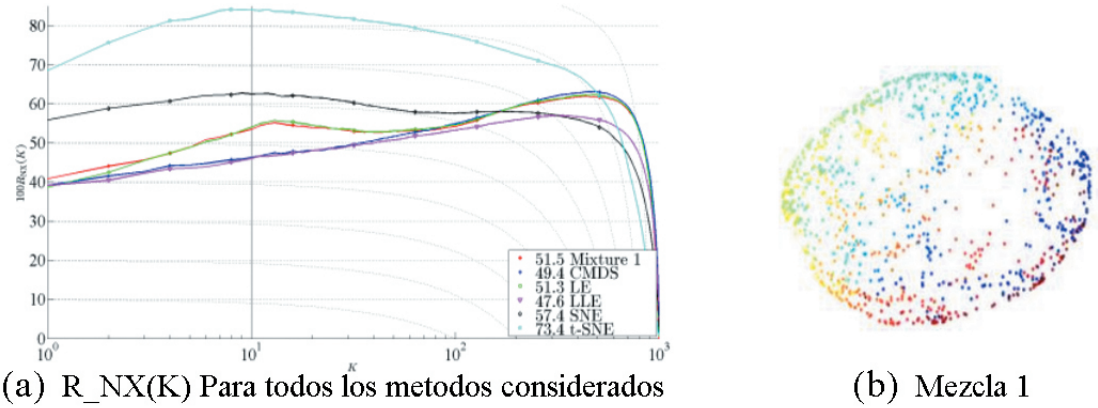
Figura 9-1.: Los efectos de los métodos de reducción de la dimensionalidad DR considerados en la estructura esférica artificial 3D. Los resultados son datos de dimensión menor representados en un espacio bidimensional.

Se observa que los métodos espectrales basados en similitudes como; LLE [88], LE [89], conservan una representación de baja dimensión que se ajusta a la morfología visual original del colector

a tratar. No obstante LLE, presenta una leve transformación horizontal dentro de su aspecto visual, el cual invita a pensar, que este método tiene una pequeña tendencia en el concepto de la preservación de los datos locales.

En métodos espectrales basados en disimilitudes como CMDS, se ve claramente la conservación máxima de su arquitectura visual (ver figura **9-1**), el cual se asume que representa los datos de tal manera, que preserva la topología global de los datos. Sin embargo dentro de su representación de menor dimensión, aunque conserva la mayoría de propiedades que caracterizan o describen mejor su estructura en términos de afinidad, no es capaz de establecer mejor aquellos factores de separabilidad intra y extra clase, tal como sucede con su homólogo PCA que se caracteriza por maximizar la varianza y tiene buenos resultados en el proceso de representación menor como factor de compresión de datos y no en términos de labores más complejas como separabilidad de clases.

Se evidencia el rendimiento de cada mezcla y para todos los métodos considerados RD. En los datos representados en baja dimensión, resultantes de cada mezcla.

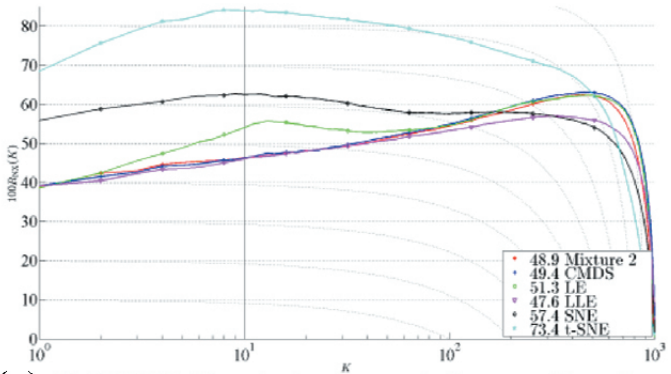


(a) R_NX(K) Para todos los metodos considerados (b) Mezcla 1

Figura 9-2.: (a) El rendimiento de la mezcla 1 y todos los métodos considerados RD. En b) se indican los datos representados en menor dimensión resultantes de la mezcla 1.

Se aprecia que en la figura (**9-2** a) el área bajo la curva de la mezcla es mayor que todos los métodos, sólo se supera con el método t-SNE. No obstante, la mezcla 1 es capaz de validar que efectivamente es un colector de topología global dado a la simetría de la evaluación del rendimiento que se puede evidenciar de forma intuitiva en su forma proyectada de su parte derecha del gráfico (**9-2** a), además de representar un balance entre las mejores cualidades de estos métodos que conservan la estructura global y además obtener propiedades, como las avanzadas y especializadas en separabilidad intra y extra clases, tales como los métodos divergentes, en el caso de t-SNE, donde se observa que la mezcla alcanza una ganancia superior a todos, gracias al porcentaje que obtiene de t-SNE, sin embargo no supera la aplicación del 100% de t-SNE. Es bien que esto suceda y que

no sea aplicado 100% este método dentro de la mezcla, dado que tal mezcla posee otras excelentes propiedades de métodos espectrales basados en similitudes, disimilitudes, que preservan la distancia, la estructura de datos y la topología de los mismos. A esto se le denomina un buen balance y controlabilidad dentro del proceso de representación de datos en dimensiones menores.



(a) R_NX(K) Para todos los metodos considerados

(b) Mezcla 2

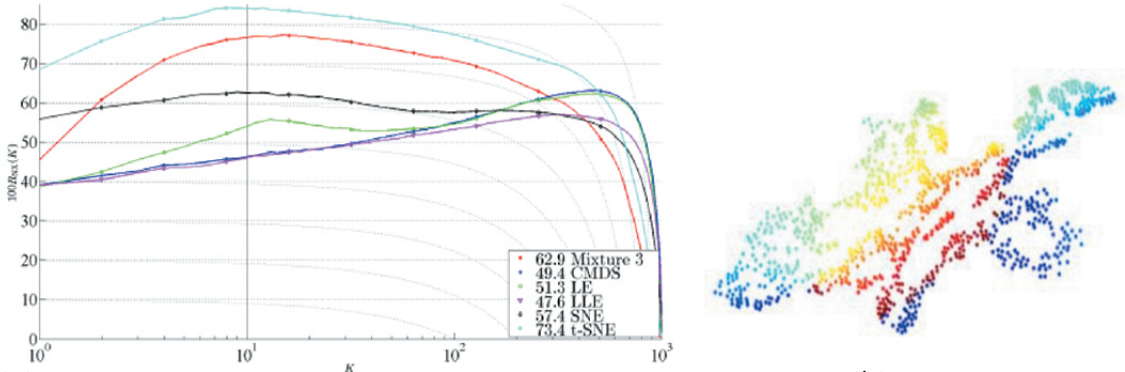
Figura 9-3.: (a) Rendimiento de la mezcla 2 y todos los métodos considerados RD. En b) se indican los datos incrustados resultantes de la mezcla 2.

Se demuestra en la figura (**9-3** a) que el área bajo la curva de la mezcla 2, representa una configuración donde mantiene un rendimiento similar, respecto la media de los rendimientos generales, con una leve tendencia a la preservación de la estructura global de los datos, tal como lo hace el método de LLE en la figura **9-1**, que posee características de transformación dentro de su forma horizontal, que proyecta una morfología visual de tipo elíptica.

Esta mezcla proyecta un balance dentro de los mejores aspectos de conservación de afinidad de la estructura original. Sin embargo hay una leve tendencia a la distribución de agrupamiento natural porque se ve claramente como el manejo del color que se encuentra embebido en cada clase, es de fácil identificación hablando en términos de agrupamiento. Para este caso, particularmente se conserva el balance de la estructura como un valor medio entre el rendimiento de todos los métodos, permitiendo conservar la controlabilidad de la dispersión de datos de la representación de menor dimensión, además de validar de igual manera, que la estructura del colector *Estructura Esférica Artificial en 3D* es global, puesto que la simetría del rendimiento en la calidad de cada método vs el rendimiento aplicado al colector se proyecta de una manera convergente o estable en el factor de 10^3, pero diverge en gran medida en el factor de 10^1.

Vale la pena resaltar el desempeño de t-SNE como método divergente, el cual presenta una superioridad en términos de poseer propiedades definidas para labores de clasificación y reducción de dimensión. A pesar de presentar dichos rendimientos, t-SNE no es capaz de controlar su factor

de transformación, el cual satisfaga una morfología que tenga factores de similitud con la representación original conservada en espacios de alta dimensión.



(a) R_NX(K) Para todos los metodos considerados (b) Mezcla 3

Figura 9-4.: (a) Realización de la mezcla 3 y todos los métodos considerados RD. En b) se indican los datos incrustados resultantes de la mezcla 3.

En concordancia con las hipótesis resultantes de las afirmaciones hechas en las mezclas 1 y 2. La mezcla 3, establece una proyección del rendimiento general que evidencia un resultado superior respecto a las dos realizaciones anteriores (Mezcla 1 (**9-2** a) y Mezcla 2 (**9-3** a)), esto se debe al porcentaje en mayor proporción de métodos divergentes que abarcan una mayor calidad dentro del colector a representar. Puesto que, las propiedades de separabilidad intra y extra clase, se establecen como un apoyo explícito a los atributos transversales de las características a explotar de los métodos seleccionados.

A pesar de lograr una transformación dramática dentro de la estructura de afinidad respecto al colector original, dicha representación se comporta muy estable en términos de agrupamientos naturales por similitud de vecindad. Esto permite proyectar la generosa área bajo la curva que proporciona los mejores rendimientos en términos de clasificación de clases. Este patrón de comportamiento, siempre está presente como factor fundamental en las medidas de rendimiento para representación en espacios de menor dimensión.

En los factores de 10^1 se evidencia un patrón de distribución casi uniforme en las medidas del rendimiento general de la mezcla versus los métodos, también se refleja un alto coeficiente en términos de ganancia en el factor 10^1 de área bajo la curva, el cual tiende a estabilizarse en el rango del factor evaluado entre 10^2 y 10^3.

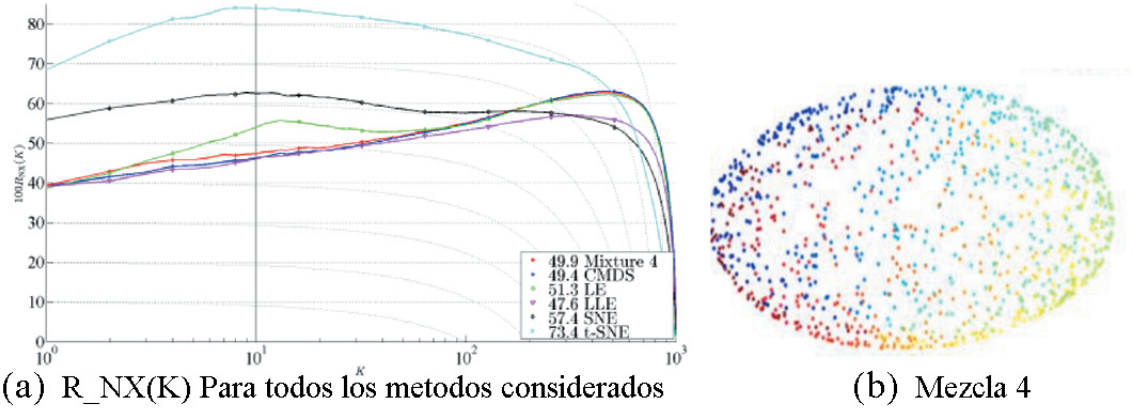


Figura 9-5.: (a) Realización de la mezcla 4 y todos los métodos considerados RD. En b) se indican los datos incrustados resultantes de la mezcla 4.

En la mezcla 4, se evidencia una similitud entre la mezcla y las representaciones del rendimiento de calidad como las proyectadas por: CMDS, LE y LLE. Esto muestra una configuración basada en métodos netamente espectrales y con muy bajos porcentajes de métodos divergentes, esto implica una desventaja al realizar un des-balanceo entre el equilibrio en los métodos seleccionados, teniendo en cuenta que la topología de la representación de los datos, claramente es de orden global.

La hipótesis nula de este proceso, concierne a la estructura global del colector *(Estructura Esférica Artificial en 3D)*. Se puede afirmar a modo de hipótesis alternativa, que este tipo de enfoques de orden fuertemente espectral, funcionan mejor para representaciones de datos de topología local. No obstante en términos visuales, obtiene una buena recompensa al establecer el grado de similitud respecto la morfología visual original, que obtiene unas leves transformaciones horizontales de tipo elíptica, que no discrepan de la representación original del colector que se encuentra, en un espacio de alta dimensión.

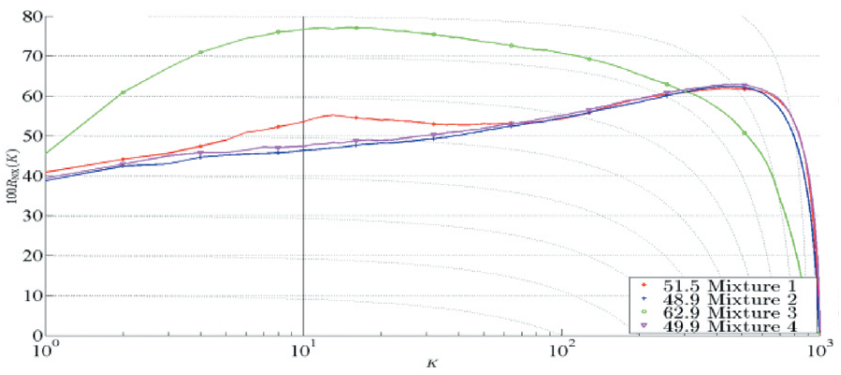


Figura 9-6.: Realización de todas las mezclas seleccionadas.

En este punto, se exponen los rendimientos generalizados de todas las mezclas aquí estudiados bajo la curva de calidad $R_{NX}(K)$, que mide el rendimiento global mediante un valor cuantitativo que se obtiene mediante la curva (AUC). La noción que se entrega de la gráfica **9-6**, corresponde a la mejor configuración que para esta investigación corresponde a la Mezcla 3. Dicha mezcla conserva el equilibrio de la representación del colector de naturaleza global, el cual establece el mayor coeficiente de área bajo la curva, por encima de aquellos rendimientos generalizados, el cual tuvo porcentajes de configuración menos idóneos para criterios de conservación de la topología estructural de los datos de tipo global, así como también para labores más complejas referente a separabilidad intra y extra clase.

El enfoque aquí propuesto, valida diversas ventajas del modelo en beneficio de la reducción de la dimensión interactiva como metodología, así como también los componentes de interacción humano ordenador especializados mediante el enfoque de INFOVIZ. Una de tantas ventajas es, permitir establecer de una manera intuitiva la naturaleza del colector o de los datos a tratar. Otra ventaja resulta en el enfoque de idoneidad de selección del algoritmo, el cual establece mediante análisis exploratorio científico mediante el canal visual, el grado asertivo respecto a la morfología resultante de los espacios de representación en menor dimensión. La curva de calidad valida lo que visualmente el usuario a modo intuitivo presume.

En los casos anteriores de las mezclas 1, 2, 3, 4 y 5 se puede afirmar que la mejor configuración podría tener coeficientes balanceados entre los métodos espectrales basados en di-similitudes en un porcentaje no mayor a 10% para mantener la controlabilidad de la dispersión en los datos sin transformaciones dramáticas que puedan cambiar la estructura final de la representación, en métodos basados en similitudes en un porcentaje no mayor a 30% para mantener la estructura global de manera articulada mediante el grado de vecindad de los nodos y aristas para la preservación de la topología y finalmente los métodos divergentes en un porcentaje restante a 60% que ayudan a maximizar la separabilidad intra y extra clase, además de poder demostrar grandes coeficientes de calidad mediante el ejercicio de la representación en un espacio de menor dimensión.

9.2. Interfaz Implementada

La figura (**9-7**) muestra la interfaz que fue diseñada para un fácil manejo donde los usuarios (ni siquiera expertos) interactúan con los métodos DR y sus posibles combinaciones de una manera intuitiva con las barras ecualizadoras y pueden visualizar sus datos en el espacio humano perceptible. La afinidad también el control deslizante es una herramienta útil para los usuarios porque pueden seguir la proximidad de los datos de alta dimensión (datos originales) en baja dimensión (datos representados en un espacio de menor dimensión) para que puedan decidir la mejor representación de sus datos.

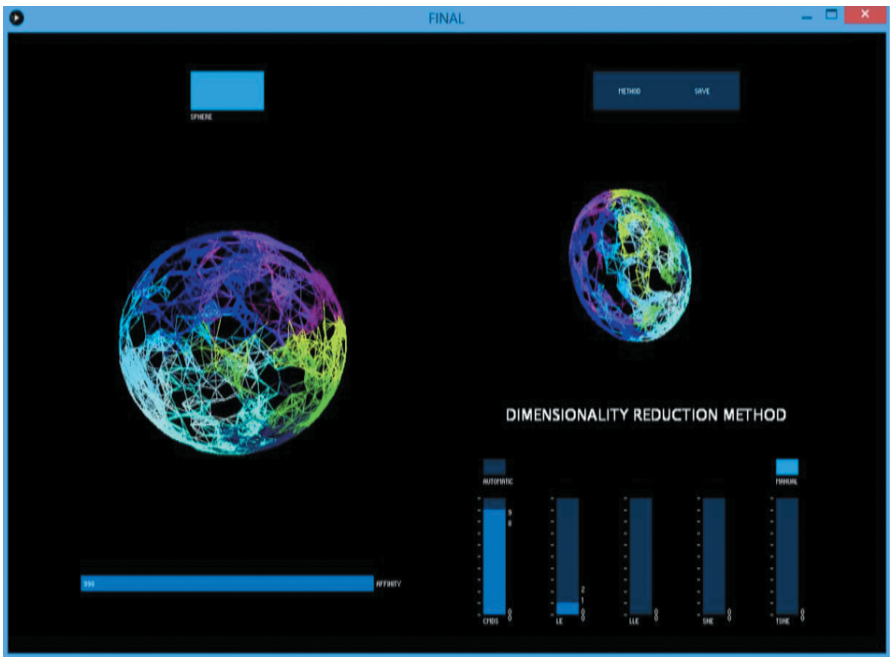


Figura 9-7.: Vista de la interfaz DataVisSim implementada en el software de procesamiento. Vídeo de muestra disponible en:

`https://youtu.be/NYh8JQ_SV4U`

Parte VI.

Observaciones Finales

10. Conclusiones y Trabajos Futuros

10.1. Conclusiones

En términos de reducción de dimensión, este trabajo presenta un nuevo enfoque interactivo de visualización de datos basado en la mezcla de resultados mediante las representaciones de dimensión menor mediante los métodos de (DR). La base de este enfoque, consiste en trazar líneas (aristas) entre los puntos de datos que presentan el valor más alto utilizando una matriz de similitud, que mide el grado de afinidad o similitud entre cada par de puntos de datos que capturan la estructura de los datos de entrada. Dicha visualización de una topología, puede ser representada por un grafo basado en datos, además de la gráfica de dispersión convencional. Las Marcas como (Puntos - "Representa la varianza de la estructura de los datos en alta dimensión") y canales como el color, proporciona una experiencia de usuario más expresiva y efectiva al utilizar también el modelo ecualizador, que proporciona la noción de interactividad al usuario, para una selección y combinación de métodos DR, mientras que proporciona información sobre la estructura de los datos originales, de tal manera que los puntos de datos representan los nodos, y una matriz de afinidad mantiene los pesos de aristas a pares.

En interacción humano ordenador, este trabajo presenta el modelo metodológico en beneficio del aporte en las representaciones de menor dimensión, que se establecen mediante el análisis exploratorio científico, dado por el modelo de ecualizador de DataVisSim y las morfologías visuales, como apoyo al aumento de las capacidades analíticas humanas, aplicado a datos masivos. Tales aportes esbozan temáticas relacionadas a: Visualización de Datos, como herramienta de explotación del canal con mayor ancho de banda en el sistema humano "la visión", Análisis Exploratorio, como herramienta probatoria en búsqueda de hipótesis, refinamiento y depuración para determinar patrones que claramente aportan a componentes relacionados, con labores de minería de datos y descubrimiento de conocimiento.

En materia de argumentos cuantitativos, la reducción de la dimensión interactiva permite establecer un nuevo paradigma algorítmico a través del modelo matemático propuesto, dado que cada modelo o configuración se re-calcula para ser representada en un espacio de menor dimensión, que satisface el problema de selección del algoritmo por idoneidad. En este punto, se aumenta las capacidades analíticas y científicas de forma intuitiva, explotando las mejores características de los algoritmos, que conllevan a la creación de N nuevos métodos, bajo la aplicación de la mezcla interactiva de (RD). Dicho proceso permite encontrar, la mejor configuración resultante para con-

servar el balance, con altos coeficientes del rendimiento de la calidad de la representación, que de otra forma no podría ser logrado, si solamente se aplican los métodos de forma básica.

10.2. Trabajo Futuro

Como trabajo futuro, se recomienda la aplicación de otros métodos de reducción de la dimensión, que se integrarán en un grafo basado en datos, con el fin de lograr una buena relación entre la preservación de la estructura de los datos y la visualización inteligible de los mismos. Se explorarán más propiedades matemáticas para diseñar los datos que mejor se aproximen, a las representaciones originales de la topología.

También se recomienda utilizar otro tipo de representaciones de bases de datos de carácter no artificial, el cual permitan evidenciar los componentes multivariados de la naturaleza de los datos masivos en alta dimensión. Esto conlleva a utilizar nuevas métricas para medir el rendimiento de los factores inmersos en la calidad de las nuevas representaciones de los datos, así como también para determinar, sí su topología es de orden global o local.

10.3. Discusiones

En esta sección se discuten los enfoques pertenecientes a la visualización usando reducción de la dimensión e interactividad, así como también el enfoque en representación visual.

Se afirma que la mayoría de sistemas de análisis visual del estado del arte, integran transversalmente el componente interactivo, el cual imposibilita la noción intuitiva que a su vez, obliga a soportar procesos muy complejos como un sistema superior, que obligatoriamente será manipulado por un experto. Tal necesidad revela 3 escenarios comunes, que subyacen de procesos de interacción para concebir susceptibilidad en ejercicios de control interactivo como: restricciones algorítmicas, selección de características y la selección de la idoneidad entre una gama de algoritmos de (RD) [143].

Autores como: Lanjing Zhang, Dominik Sacha y Michael Sedlmair, defienden (La interacción visual con la reducción de la dimensionalidad en complejidades de tiempos de ejecución) [144]. El desarrollo de tal afirmación, evidencia los esfuerzos de centrar las investigaciones en implementaciones ingenieriles específicas de sistemas para el análisis visual integrando RD, de manera que resultan en el análisis de las formas, en beneficio de los costos computacionales, pero carecen del factor intuitivo del diseño visual para lograr, la anhelada expresividad y efectividad con los canales

y marcas mas efectivos.

Otros autores como Vince D. Calhoun, discrepan en aquellas representaciones de baja dimensión en beneficio de la exploración de la separabilidad de clases, porque carecen del criterio que identifica las capacidades perceptivas en los humanos. En este punto, se empieza a entender la importancia de los factores de expresividad y efectividad como aumento a las capacidades analíticas, en labores de exploración de datos [145].

Autores como Anant Madabhushi, poseen una fuerte inclinación para ejecutar los métodos de RD, en beneficio de un conocimiento previo. Dicha afirmación establece siempre el conocimiento a priori de un sistema superior, el cual es una gran desventaja al requerir de un experto.

Algunos autores como Xiaoru Yuan, Donghao Ren han avanzado en el enfoque con más resultados positivos y aceptados por la comunidad de Reducción de la Dimensión, HCI e INFOVIZ. el enfoque de idoneidad en selección del algoritmo, desarrolla especial interés, en todos los movimientos intermedios que específicamente focalizan esfuerzos en procesos de exploración mediante la interpolación general, dicho enfoque espera que los algoritmos de RD no convexos, sean manipulados por eventos interactivos bajo cambios de parámetros del lenguaje de diseño, en el espacio de entrada para que puedan tener una correspondencia en su proceso de interactividad, que permita manejar los parámetros en función de un factor humano, y las salidas de la representación [147]. Sin embargo solucionan solo un problema que es la interacción, pero dejan de lado la expresividad y efectividad mediante los canales y marcas más efectivas tal como lo expresa Tammara Munzner.

De acuerdo a los enfoques de visualización usando reducción de la dimensión e interactividad mencionados, se afirma que la codificación visual, apoya directamente procesos que aumentan las capacidades analíticas humanas, en beneficio de establecer aquellas morfologías visuales que tengan un alto grado de correspondencia en métricas de efectividad y expresividad en cuanto a Canales (Color, Movimiento, Forma, Región Espacial.) y Marcas (Primitivas Geométricas). Tal componente se desarrolló en el modelo DataVisSim, aquí expuesto. De esta manera se complementa el estado del arte mediante una nueva propuesta que lo integra todo, de una manera intuitiva, sin necesidad de utilizar procesos de sistemas superiores que obligatoriamente, necesitan de un experto, en temáticas tan especializadas como Reducción de la Dimensión. Como último componente se introduce en esta investigación, el análisis exploratorio para el descubrimiento científico, permitiendo refinar, depurar o ampliar el modelo matemático aquí expuesto (Modelo de Ecualizador), para tener un entendimiento del fenómeno, que estudia la manera en que se afectan las salidas determinísticas de tales modelos a través de los eventos mediante el paso de parámetros.

Parte VII.

Bibliografía

Bibliography

[1] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[2] Cisco, "Internet será cuatro veces más grande en 2016," *Neurocomputing*, vol. 3, pp. 23–45, 2011.

[3] P. Russom *et al.*, "Big data analytics," *TDWI best practices report, fourth quarter*, vol. 19, p. 40, 2011.

[4] C. Dai, "Big data: The data velocity discussion," *thinking.netezza.com*, vol. 1, pp. 1–9, 2014.

[5] C. Ballard, K. Foster, A. Frenkiel, B. Gedik, M. P. Koranda, S. Nathan, D. Rajan, R. Rea, M. Spicer, B. Williams *et al.*, "Ibm infosphere streams: Assembling continuous insight in the information revolution," 2012.

[6] S. Sunil, "Not your type? big data matchmaker on five data types you need to explore today," *dataversity.net*, vol. 2, pp. 3–7, 2015.

[7] hadoop.apache.org, "Aprenda más acerca de apache hadoop," *hadoop.apache.org*, vol. 1, pp. 2–3, 2015.

[8] A. J. Anaya-Isaza, D. H. Peluffo-Ordoñez, J. C. Alvarado-Pérez, J. Ivan-Rios, J. A. Castro-Silva, P. D. Rosero-Montalvo, D. F. Peña-Unigarro, and A. C. Umaquinga-Criollo, "Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: Lda versus pca."

[9] D. H. Peluffo-Ordónez, J. C. Alvarado-Pérez, J. A. Lee, M. Verleysen *et al.*, "Geometrical homotopy for data visualization," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015)*, 2015.

[10] E. Tufte, "The visual display of quantitative information paperback," *Graphics Press*, vol. 2, pp. 10–11, 2011.

[11] Y. Wei, S. Liu, J. Sun, L. Cui, L. Pan, and L. Wu, "Big datasets for research: A survey on flagship conferences," in *Big Data (BigData Congress), 2016 IEEE International Congress on*. IEEE, 2016, pp. 394–401.

[12] J. Salazar-Castro, Y. Rosas-Narváez, A. Pantoja, J. C. Alvarado-Pérez, and D. H. Peluffo-Ordóñez, "Interactive interface for efficient data visualization via a geometric approach," in *Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on*. IEEE, 2015, pp. 1–6.

[13] M. D. O. Morales, L. J. Aguilar, and L. M. G. Marín, "Los desafíos del marketing en la era del big data," *e-Ciencias de la Información*, vol. 6, no. 1, pp. 1–31, 2015.

[14] Y. Demchenko, E. Gruengard, and S. Klous, "Instructional model for building effective big data curricula for online and campus education," in *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on*. IEEE, 2014, pp. 935–941.

[15] P. D. C. de Almeida and J. Bernardino, "Big data open source platforms," in *Big Data (BigData Congress), 2015 IEEE International Congress on*. IEEE, 2015, pp. 268–275.

[16] Y. M. Seo and W. K. Kim, "Information visualization process for spatial big data," *Journal of Korea Spatial Information Society*, vol. 23, no. 6, pp. 109–116, 2015.

[17] P. Chaudhary and V. K. Yadav, "A survey on security issues and the existing solutions in big data," *International Journal of Computer Applications*, vol. 162, no. 1, 2017.

[18] J. C. Riquelme Santos, R. Ruiz, and K. Gilbert, "Mineria de datos: Conceptos y tendencias," *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial*, vol. 10, no. 29, pp. 11–18, 2006.

[19] C. Logreira, "Minería de datos y su incidencia en la toma de decisiones empresariales en el contexto de crm," *Ingeniería solidaria*, vol. 7, no. 13, pp. 68–71, 2011.

[20] M. C. Hao, U. Dayal, D. A. Keim, and T. Schreck, "Multi-resolution techniques for visual exploration of large time-series data," in *EUROVIS 2007*, 2007, pp. 27–34.

[21] P. C. Wong, "Visual data mining," *IEEE Computer Graphics and Applications*, vol. 19, no. 5, pp. 20–21, 1999.

[22] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Information Visualization, 2006. IV 2006. Tenth International Conference on*. IEEE, 2006, pp. 9–16.

[23] C. Ahlberg and E. Wistrand, "Ivee: An information visualization and exploration environment," in *Information Visualization, 1995. Proceedings.* IEEE, 1995, pp. 66–73.

[24] A. Kerren, A. Ebert, and J. Meyer, *Human-centered visualization environments*. Springer-Verlag Berlin Heidelberg, 2007.

[25] L. Manovich, "What is visualisation?" *Visual Studies*, vol. 26, no. 1, pp. 36–49, 2011.

[26] ——, "Visualization methods for media studies," *Oxford Handbook of Sound and Image in Digital Media*, pp. 253–78, 2014.

[27] H. Lam, M. Tory, and T. Munzner, "Bridging from goals to tasks with design study analysis reports," *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[28] L. Manovich, "What is visualization?" *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture*, vol. 2, no. 1, 2010.

[29] A. A. Salah, L. Manovich, A. A. Salah, and J. Chow, "Combining cultural analytics and networks analysis: Studying a social network site with user-generated content," *Journal of Broadcasting & Electronic Media*, vol. 57, no. 3, pp. 409–426, 2013.

[30] M. Brehmer, B. Lee, B. Bach, N. H. Riche, and T. Munzner, "Timelines revisited: A design space and considerations for expressive storytelling," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 9, pp. 2151–2164, 2017.

[31] A. Batch and N. Elmqvist, "The interactive visualization gap in initial exploratory data analysis," *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[32] E. Bendoly, "Fit, bias, and enacted sensemaking in data visualization: frameworks for continuous development in operations and supply chain management analytics," *Journal of Business Logistics*, vol. 37, no. 1, pp. 6–17, 2016.

[33] C. Moten III, H. Newton, and L. Jackson, "Trac innovative visualization techniques," TRAC-Monterey Monterey United States, Tech. Rep., 2016.

[34] M. Brehmer, J. Ng, K. Tate, and T. Munzner, "Matches, mismatches, and methods: multiple-view workflows for energy portfolio analysis," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 449–458, 2016.

[35] A. Crisan, J. L. Gardy, and T. Munzner, "On regulatory and organizational constraints in visualization design and evaluation," *arXiv preprint arXiv:1610.10056*, 2016.

[36] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller, "Visual parameter space analysis: A conceptual framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2161–2170, 2014.

[37] T. Munzner, "Keynote speaker: Visualization analysis and design," in *Pacific Visualization Symposium (PacificVis), 2016 IEEE*. IEEE, 2016, pp. xiii–xiii.

[38] A. Crisan, J. L. Gardy, and T. Munzner, "On regulatory and organizational constraints in visualization design and evaluation," *arXiv preprint arXiv:1610.10056*, 2016.

[39] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 519–528, 2016.

[40] G. G. Méndez, M. A. Nacenta, and S. Vandenheste, "ivolver: Interactive visual language for visualization extraction and reconstruction," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 4073–4085.

[41] K. Gupta, S. Sampat, M. Sharma, and V. Rajamanickam, "Visualization of election data: Using interaction design and visual discovery for communicating complex insights," *JeDEM-eJournal of eDemocracy and Open Government*, vol. 8, no. 2, pp. 59–86, 2016.

[42] A. Handler, S. L. Blodgett, and B. O'Connor, "Visualizing textual models with in-text and word-as-pixel highlighting," *arXiv preprint arXiv:1606.06352*, 2016.

[43] M. L. Young, A. Hermida, and J. Fulda, "What makes for great data journalism? a content analysis of data journalism awards finalists 2012–2015," *Journalism Practice*, pp. 1–21, 2017.

[44] P. Mi, "Gpu based methods for interactive information visualization of big data," Ph.D. dissertation, Virginia Tech, 2016.

[45] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[46] T. Munzner, "Visualization analysis and design for business intelligence."

[47] U. Ashish, A. R. Shankar, S. Pai, and S. Anguru, "Novel dpi technique to decode secret information for network security."

[48] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, D. A. Keim, J. Krause, A. Perer, E. Bertini, T. N. Dang *et al.*, "Ieee visual analytics science and technology conference."

[49] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Toward a deeper understanding of visualization through keyword analysis," *arXiv preprint arXiv:1408.3297*, 2014.

[50] K. Azoumana, "Análisis de la deserción estudiantil en la universidad simón bolívar, facultad ingeniería de sistemas, con técnicas de minería de datos," *Revista Pensamiento Americano*, vol. 6, no. 10, 2014.

[51] P. Santana Mansilla, R. Costaguta, and D. Missio, "Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos," *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 17, no. 53, 2014.

[52] ——, "Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos," *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 17, no. 53, 2014.

[53] A. Castro, E. Sifuentes, S. González, and L. H. Rascón, "Uso de minería de datos en el manejo de información geográfica," *Información tecnológica*, vol. 25, no. 5, pp. 95–102, 2014.

[54] A. A. R. Gómez and A. V. Ibáñez, "Minería de datos aplicada a la demanda del transporte aéreo en ocaña, norte de santander," *Revista Tecnura*, vol. 19, no. 45, pp. 101–114, 2015.

[55] K. B. Eckert and R. Suénaga, "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos," *Formación universitaria*, vol. 8, no. 5, pp. 03–12, 2015.

[56] J. Izquierdo, E. Campbell, I. Montalvo, and R. Pérez-García, "Combinación multi-agente de algoritmos evolutivos y minería de datos para mejorar la búsqueda en problemas de optimización del mundo real," in *Congresso de Métodos Numéricos em Engenharia*, 2015.

[57] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[58] D. T. Larose, *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[59] A. Holzinger, M. Dehmer, and I. Jurisica, "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions," *BMC bioinformatics*, vol. 15, no. 6, p. I1, 2014.

[60] G. Shmueli and K. C. Lichtendahl Jr, *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons, 2017.

[61] A. Holzinger and I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in *Interactive knowledge discovery and data mining in biomedical informatics*. Springer, 2014, pp. 1–18.

[62] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning analytics*. Springer, 2014, pp. 61–75.

[63] I. Borg and P. J. Groenen, "Multidimensional scaling theory and applications," *Springer*, vol. 28, no. 3, pp. 317–320, 1998.

[64] L. Wolf and S. Bileschi, "Combining variable selection with dimensionality reduction," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 801–806.

[65] D. H. Peluffo, J. A. Lee, M. Verleysen, J. L. Rodrıguez, and G. Castellanos-Domınguez, "Unsupervised relevance analysis for feature extraction and selection."

[66] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1855–1887, 2005.

[67] F. D. C. Sotelo J. Rodríguez, Diego Hernán Peluffo Ordo˜nez and G. Castellanos-Domínguez, "Unsupervised feature relevance analysis applied to improve ecg heartbeat clustering." *Journal of Computer Methods and Programs in Biomedicine*, vol. 2, no. Nov, pp. 250–261, 2012., 2012.

[68] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[69] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[70] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proceedings of the twenty first international conference on Machine learning*. ACM, 2004, p. 47.

[71] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of kullback leibler divergences as cost functions in dimensionality reduction based on similarity preservation," *Neurocomputing*, vol. 112, pp. 92–108, 2013.

[72] D. Pimentel, M. Cataldi, and G. Muñiz, "De la visualización a la sensorización de información," *Blucher Design Proceedings*, vol. 1, no. 7, pp. 129–133, 2013.

[73] C. E. M. Echeverry, M. L. Trujillo, and M. H. M. Salazar, "Minería de datos en gestión del conocimiento de pymes de colombia," *Revista Virtual Universidad Católica del Norte*, no. 50, pp. 224–237, 2017.

[74] R. Ohannessian, T. Bénet, L. Argaud, C. Guérin, C. Guichon, V. Piriou, T. Rimmelé, R. Girard, S. Gerbier-Colomban, and P. Vanhems, "Heat map for data visualization in infection control epidemiology: An application describing the relationship between hospital-acquired infections, simplified acute physiological score ii, and length of stay in adult intensive care units," *American Journal of Infection Control*, 2017.

[75] M. Angelini, T. Catarci, M. Mecella, and G. Santucci, "The visual side of the data," in *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*. Springer, 2018, pp. 3–25.

[76] A. Inselberg and L. G. Anthopoulos, "Visual analytics for high dimensional data: Very late added paper," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 1683–1687.

[77] M. Lu, C. Lai, T. Ye, J. Liang, and X. Yuan, "Visual analysis of multiple route choices based on general gps trajectories," *IEEE Transactions on Big Data*, 2017.

[78] T. A. Snijders, "Stochastic actor-oriented models for network dynamics," *Annual Review of Statistics and Its Application*, vol. 4, pp. 343–363, 2017.

[79] D. A. Keim and H.-P. Kriegel, "Visdb: Database exploration using multidimensional visualization," *IEEE Computer Graphics and Applications*, vol. 14, no. 5, pp. 40–49, 1994.

[80] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[81] D. A. Keim and H.-P. Kriegel, "Visualization techniques for mining large databases: A comparison," *IEEE Transactions on knowledge and data engineering*, vol. 8, no. 6, pp. 923–938, 1996.

[82] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery," in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*. ACM, 2009, pp. 12–20.

[83] D. H. Peluffo-Ordónez, J. A. Lee, and M. Verleysen, "Short review of dimensionality reduction methods based on stochastic neighbour embedding," in *Advances in Self-Organizing Maps and Learning Vector Quantization*. Springer, 2014, pp. 65–74.

[84] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[85] J. B. Tenenbaum, "Mapping a manifold of perceptual observations," in *Advances in neural information processing systems*, 1998, pp. 682–688.

[86] S. Weiren and W. Kai, "Floyd algorithm for the shortest path planning of mobile robot [j]," *Chinese Journal of Scientific Instrument*, vol. 10, pp. 2088–2092, 2009.

[87] S. Skiena, "Dijkstra's algorithm," *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley*, pp. 225–227, 1990.

[88] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[89] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[90] ——, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[91] C. T. Baker, "The numerical treatment of integral equations," 1977.

[92] Y. Bengio, J.-f. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," in *Advances in neural information processing systems*, 2004, pp. 177–184.

[93] Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, and N. Le Roux, *Spectral clustering and kernel PCA are learning eigenfunctions*. CIRANO, 2003, vol. 1239.

[94] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[95] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 21, pp. 7426–7431, 2005.

[96] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[97] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[98] X. Y. Stella and J. Shi, *Multiclass spectral clustering*, 2003.

[99] A. Hyvarinen, "Independent component analysis: algorithms and applications," *Neural Network*, vol. 13, pp. 411–430, 2000.

[100] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM review*, vol. 35, no. 4, pp. 551–566, 1993.

[101] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. Van Der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems*, 2001, vol. 27, no. 4.

[102] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[103] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[104] G. Young and A. S. Householder, "Discussion of a set of points in terms of their mutual distances," *Psychometrika*, vol. 3, no. 1, pp. 19–22, 1938.

[105] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[106] J. de Leeuw and W. Heiser, "13 theory of multidimensional scaling," *Handbook of statistics*, vol. 2, pp. 285–316, 1982.

[107] R. Abrams, N. Galanter, D. Harness, and C. Parker, "Machine learning for the classification of toxicological endpoints," 2017.

[108] B. Jeon, J. Jung, B. D. Youn, Y. Kim, and Y.-C. Bae, *Statistical approach to diagnostic rules for various malfunctions of journal bearing system using Fisher discriminant analysis*, 2014.

[109] R. A. Damon Jr and W. R. Harvey, "Experimental design, anova, and regression," no. 311.2 D163, 1987.

[110] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[111] ——, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[112] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods." in *Fgr*, vol. 2, 2002, p. 215.

[113] J. Yang, A. F. Frangi, J.-y. Yang, D. Zhang, and Z. Jin, "Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[114] H. H. Harman, *Modern factor analysis*. University of Chicago Press, 1976.

[115] E. E. Gorenstein, C. A. Mammato, and J. M. Sandy, "Performance of inattentive-overactive children on selected measures of prefrontal-type function," *Journal of clinical psychology*, vol. 45, no. 4, pp. 619–632, 1989.

[116] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE transactions on image processing*, vol. 11, no. 2, pp. 146–158, 2002.

[117] G. Van de Wouwer, P. Scheunders, and D. Van Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE transactions on image processing*, vol. 8, no. 4, pp. 592–598, 1999.

[118] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[119] Z. Yang, I. King, Z. Xu, and E. Oja, "Heavy-tailed symmetric stochastic neighbor embedding," in *Advances in neural information processing systems*, 2009, pp. 2169–2177.

[120] D. K. Agrafiotis, H. Xu, F. Zhu, D. Bandyopadhyay, and P. Liu, "Stochastic proximity embedding: methods and applications," *Molecular informatics*, vol. 29, no. 11, pp. 758–770, 2010.

[121] D. K. Agrafiotis, "Stochastic proximity embedding," *Journal of computational chemistry*, vol. 24, no. 10, pp. 1215–1221, 2003.

[122] L. Xu, "Least mean square error reconstruction principle for self-organizing neural-nets," *Neural networks*, vol. 6, no. 5, pp. 627–648, 1993.

[123] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*, 1974.

[124] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[125] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[126] P. Demartines and J. Hérault, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *IEEE Transactions on neural networks*, vol. 8, no. 1, pp. 148–154, 1997.

[127] J. Hérault, C. Jausions-Picaud, and A. Guérin-Dugué, "Curvilinear component analysis for high-dimensional data representation: I. theoretical aspects and practical use in the presence of noise," *Engineering Applications of Bio-Inspired Artificial Neural Networks*, pp. 625–634, 1999.

[128] J. Venna and S. Kaski, "Visualizing gene interaction graphs with local multidimensional scaling." in *ESANN*, vol. 6, 2006, pp. 557–562.

[129] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[130] J. A. Lee, A. Lendasse, M. Verleysen *et al.*, "Curvilinear distance analysis versus isomap." in *ESANN*, vol. 2, 2002, pp. 185–192.

[131] J. A. Lee, C. Archambeau, M. Verleysen *et al.*, "Locally linear embedding versus isotop." in *ESANN*, 2003, pp. 527–534.

[132] M. Sabin and R. Gray, "Global convergence and empirical consistency of the generalized lloyd algorithm," *IEEE Transactions on information theory*, vol. 32, no. 2, pp. 148–155, 1986.

[133] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[134] C. M. Bishop, M. Svensén, and C. K. Williams, "Gtm: The generative topographic mapping," *Neural computation*, vol. 10, no. 1, pp. 215–234, 1998.

[135] ——, "Gtm: A principled alternative to the self-organizing map," in *Advances in neural information processing systems*, 1997, pp. 354–360.

[136] I. D. Blanco, A. A. C. Vega, and A. B. D. González, "Correlation visualization of high dimensional data using topographic maps," in *International Conference on Artificial Neural Networks*. Springer, 2002, pp. 1005–1010.

[137] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[138] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers*, vol. 100, no. 5, pp. 401–409, 1969.

[139] J. Mao and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE transactions on neural networks*, vol. 6, no. 2, pp. 296–317, 1995.

[140] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[141] J. A. Lee and M. Verleysen, "Nonlinear dimensionality reduction of data manifolds with essential loops," *Neurocomputing*, vol. 67, pp. 29–53, 2005.

[142] M. H. Law, N. Zhang, and A. K. Jain, "Nonlinear manifold learning for data stream," in *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM, 2004, pp. 33–44.

[143] T. Siameh, "Graph analytics methods in feature engineering," 2017.

[144] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 241–250, 2017.

[145] C. A. Tamminga, G. D. Pearlson, A. D. Stan, R. D. Gibbons, J. Padmanabhan, M. Keshavan, and B. A. Clementz, "Strategies for advancing disease definition using biomarkers and genetics: The bipolar and schizophrenia network for intermediate phenotypes," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 2, no. 1, pp. 20–27, 2017.

[146] H. Kim, J. Choo, C. Lee, H. Lee, C. K. Reddy, and H. Park, "Pive: Per-iteration visualization environment for real-time interactions with dimension reduction and clustering." in *AAAI*, 2017, pp. 1001–1009.

[147] I. Dıaz, A. A. Cuadrado, D. Pérez, F. J. Garcıa, and M. Verleysen, "Interactive dimensionality reduction for visual analytics," in *Proceedings of the 22th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014)*, 2014, pp. 183–188.

[148] S. E. Viswanath, P. Tiwari, G. Lee, and A. Madabhushi, "Dimensionality reduction-based fusion approaches for imaging and non-imaging biomedical data: concepts, workflow, and use-cases," *BMC medical imaging*, vol. 17, no. 1, p. 2, 2017.

[149] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.

[150] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation," *Neurocomputing*, 2013.

[151] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[152] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 313–319.

[153] I. Borg, *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[154] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[155] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, 2002, pp. 833–840.

[156] J. de Leeuw and W. Heiser, "13 theory of multidimensional scaling," *Handbook of statistics*, vol. 2, pp. 285–316, 1982.

[157] M.-Y. Cho and H. T. Thom, "Fault diagnosis for distribution networks using enhanced support vector machine classifier with classical multidimensional scaling." *Journal of Electrical Systems*, vol. 13, no. 3, 2017.

[158] Y. Bahroun and A. Soltoggio, "Online representation learning with single and multi-layer hebbian networks for image classification," in *International Conference on Artificial Neural Networks*. Springer, 2017, pp. 354–363.

[159] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[160] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[161] R. Ji, H. Liu, L. Cao, D. Liu, Y. Wu, and F. Huang, "Toward optimal manifold hashing via discrete locally linear embedding," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5411–5420, 2017.

[162] X. Liu, D. Tosun, M. W. Weiner, N. Schuff, A. D. N. Initiative *et al.*, "Locally linear embedding (lle) for mri based alzheimer's disease classification," *Neuroimage*, vol. 83, pp. 148–157, 2013.

[163] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[164] ——, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[165] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[166] M. A. Carreira-Perpinán, "The elastic embedding algorithm for dimensionality reduction." in *ICML*, vol. 10, 2010, pp. 167–174.

[167] D. H. Peluffo-Ordónez, J. A. Lee, and M. Verleysen, "Short review of dimensionality reduction methods based on stochastic neighbour embedding," in *Advances in Self-Organizing Maps and Learning Vector Quantization*. Springer, 2014, pp. 65–74.

[168] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[169] W. Dai and P. Hu, "Research on personalized behaviors recommendation system based on cloud computing," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 1480–1486, 2013.

[170] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010.

[171] I. Díaz Blanco, A. A. Cuadrado Vega, D. Pérez López, F. J. García Fernández, and M. Verleysen, "Interactive dimensionality reduction for visual analytics," in *ESANN 2014, 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, April, 23, 24, 25, 2014: proceedings*. ESANN, 2014.

[172] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery," in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*. ACM, 2009, pp. 12–20.

[173] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[174] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[175] M. Lichman, "Uci machine learning repository," 2013.

[176] L. Roszkowiak, A. Korzynska, J. Zak, D. Pijanowska, Z. Swiderska-chadaj, and T. Markiewicz, "Survey: interpolation methods for whole slide image processing," *Journal of microscopy*, vol. 265, no. 2, pp. 148–158, 2017.

[177] G. Scheuermann, X. Tricoche, and H. Hagen, "C1-interpolation for vector field topology visualization," in *Visualization'99. Proceedings*. IEEE, 1999, pp. 271–533.

[178] C.-C. Chang, T.-S. Nguyen, and Y. Liu, "A reversible data hiding scheme for image interpolation based on reference matrix," in *Biometrics and Forensics (IWBF), 2017 5th International Workshop on*. IEEE, 2017, pp. 1–6.

[179] X. Zhang, Z. Sun, Z. Tang, C. Yu, and X. Wang, "High capacity data hiding based on interpolated image," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9195–9218, 2017.

[180] V. Sonneville, O. Brüls, and O. A. Bauchau, "Interpolation schemes for geometrically exact beams: A motion approach," *International Journal for Numerical Methods in Engineering*, 2017.

[181] S.-Y. Baek and K. Lee, "An isometric shape interpolation method on mesh models," *Korean Journal of Computational Design and Engineering*, vol. 19, no. 2, pp. 119–128, 2014.

[182] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[183] G. E. Hinton and T. J. Sejnowski, *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[184] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7, pp. 1431–1443, 2009.

Parte VIII.

Anexos

A. Producción Intelectual

Con el desarrollo de este trabajo, se logró los siguientes productos académicos:

1. **A. J. Anaya-Isaza**, D. H. Peluffo-Ordóñez, J. C. Alvarado-Pérez, J. Ivan-Rios, J. A. Castro-Silva, P. D. Rosero-Montalvo, D. F. Peña-Unigarro, J. A. Salazar-Castro, A. C. Umaquinga-Criollo. *" Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: LDA versus PCA. "*. In: 2016 International Conference on Information Systems and Computer Science (INCISCOS). Quito, Ecuador. Available from: http://ingenieria.ute.edu.ec/conferencias/index.php/inciscos/2016/paper/view/31

2. D. H. Peluffo-Ordóñez, M. A. Becerra, A. E. Castro-Ospina, X. Blanco-Valencia, J. C. Alvarado-Pérez, R. Therón, **A. Anaya-Isaza**. *"On the Relationship Between Dimensionality Reduction and Spectral Clustering from a Kernel Viewpoint"*. In: 13th International Conference on Distributed Computing and Artificial Intelligence, Advances in Intelligent Systems and Computing. DCAI 2016. Seville, Spain. ISSN:2194-5357. Available from: http://link.springer.com/chapter/10.1007/978-3-319-40162-1_28

3. P. D. Rosero-Montalvo, P. Diaz, J. A. Salazar-Castro, D. F. Peña-Unigarro, **A. J. Anaya-Isaza**, J. C. Alvarado-Pérez, R. Therón, D. H. Peluffo-Ordóñez. *" Interactive data visualization using dimensionality reduction and similarity-based representations "*. In: 2016 XXI IberoAmerican Congress on Pattern Recognition (CIARP). Cartagena, Colombia. Available from: http://link.springer.com/chapter/10.1007/978-3-319-52277-7_41

4. P. D. Rosero-Montalvo, P. Diaz, J. A. Salazar-Castro, D. F. Peña-Unigarro, **A. J. Anaya-Isaza**, J. C. Alvarado-Pérez, R. Therón, D. H. Peluffo-Ordóñez. *" Interactive data visualization using dimensionality reduction and similarity-based representations "*. In: 2016 XXI IberoAmerican Congress on Pattern Recognition (CIARP). Lima, Perú. Available from: http://link.springer.com/chapter/10.1007/978-3-319-52277-7_41

5. A. C. Umaquinga-Criollo, D. H. Peluffo-Ordóñez, M. V. Cabrera-Álvarez, J. C. Alvarado-Pérez, **A. J. Anaya-Isaza**. *" Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva "*. In: 2016 Jornadas Internacionales FICA. Ibarra, Ecuador. Available from:

http://diegopeluffo.com/publicaciones/2016_JornadasFica_BigData.pdf

6. **A. J. Anaya-Isaza**, D. H. Peluffo-Ordóñez, J. Ivan-Rios, J. A. Castro-Silva, D. A. Carvajal-Ruiz, L. H. Espinoza-Llanos. *" Sistema de Riego Basado En La Internet De Las Cosas (IoT) "*. In: 2016 Jornadas Internacionales FICA. Ibarra, Ecuador. Available from: http://diegopeluffo.com/publicaciones/2016_JornadasFica_IOT.pdf

B. Material Suplementario

Aquí se exponen algunos elementos adicionales, como el sitio web de la tesis (ver B.1), con un breve resumen para dar una noción rápida a cualquier investigador o persona que realice investigaciones similares o concernientes al tema de reducción de la dimensión e interactividad. También se expone las citas registradas hasta el momento de acuerdo al portal de Google Académico con su post Google Citations (ver B.2).

B.1. Sitio Web de la Tesis

En esta sección se muestran los pantallazos del sitio, así como también el link en descripción para la página que permite ver la tesis en su versión final:
`https://sites.google.com/view/tesisandresanaya`

Figura B-1.: Pantallazo del front de la página web.

B.2. Google Citations

En esta sección se muestra el link y los pantallazos de las citaciones hasta el año 2017 a corte de diciembre:
`https://scholar.google.es/citations?user=o2MNW98AAAAJ&hl=es`

Andres Javier Anaya Isaza

SEGUIR

Profesor de Data Science Fundamental, Data Mining, Machine Learning, Universidad Surcolombiana
Dirección de correo verificada de usco.edu.co
Dimensionality Reduction Machine Learning Data Mining

CREAR MI PROPIO PERFIL

TÍTULO	CITADO POR	AÑO
Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction DF Peña-ünigarro, JA Salazar-Castro, DH Peluffo-Ordóñez, ... Signal Processing, Images and Artificial Vision (STSIVA), 2016 XXI Symposium ...	3	2016
Interactive data visualization using dimensionality reduction and similarity-based representations P Rosero-Montalvo, P Diaz, JA Salazar-Castro, DF Peña-Unigarro, ... Iberoamerican Congress on Pattern Recognition, 334-342	1	2016
On the Relationship Between Dimensionality Reduction and Spectral Clustering from a Kernel Viewpoint AAI D. H. Peluffo-Ordóñez, M. A. Becerra, A. E. Castro-Ospina, X. Blanco ... Advances in Intelligent Systems and Computing 474 (2194-5357), 7	1	2016
Sistema de Riego Basado En La Internet De Las Cosas (IoT) AJ Anaya-Isaza, DH Peluffo-Ordoñez, J Ivan-Rios, JA Castro-Silva, ...	1	
Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva Proposal for visual analysis of Big Data using interactive dimensionalit... AC Umaquinga-Criollo, DH Peluffo-Ordóñez, PD Rosero-Montalvo, ...		
Propuesta de análisis visual de datos en Big Data usando reducción de dimensión interactiva Visual Big Data analysis proposal using interactive dimensionality red... AC Umaquinga-Criollo, DH Peluffo-Ordóñez, MV Cabrera-Álvarez, ...		
Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: LDA versus PCA AJ Anaya-Isaza, DH Peluffo-Ordoñez, JC Alvarado-Pérez, J Ivan-Rios, ...		

Citado por

	Total	Desde 2012
Citas	6	6
Índice h	1	1
Índice i10	0	0

Coautores


Diego Hernán Peluffo-Ordóñez
Professor at Universidad Técnic...


Juan Carlos Alvarado-Pérez
Profesor Corporación Autónoma ...